UNIVERSIDAD MIGUEL HERNÁNDEZ

DEPARTAMENTO DE PATOLOGÍA Y CIRUGÍA

VALIDACIÓN DE EuroSCORE II EN UN CENTRO DE MEDIO VOLUMEN

TESIS DOCTORAL

Antonio José García Valentín

Alicante, 2015

DIRECTORES

Dr. D. Antonio F. Compañ Rosique

Dr. D. Carlos A. Mestres Lucio

UNIVERSITAS
Miguel
Hernández

UNIVERSITAS
Miguel
Hernández

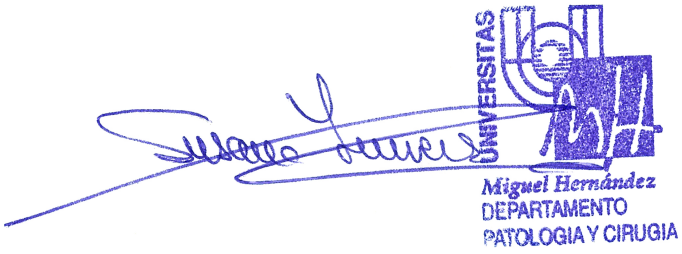
UNIVERSITAS
Miguel
Hernández

Dª María Susana Jiménez Moreno, Directora del Departamento de Patología y Cirugía de la Facultad de Medicina de la Universidad Miguel Hernández de Elche

CERTIFICA:

Que D. Antonio García Valentín ha realizado bajo la coordinación de este Departamento su memoria de tesis doctoral titulada "VALIDACIÓN DE EuroSCORE II EN UN CENTRO DE MEDIO VOLUMEN" cumpliendo todos los objetivos previstos, finalizando su trabajo de forma satisfactoria para su defensa pública y capacitándole para optar al Grado de Doctor.

Lo que certifico en San Joan d´Alacant a 10 de Diciembre de 2015.



Fdo. Dª María Susana Jiménez Moreno

UNIVERSITAS
Miguel
Hernández

D. Antonio F. Compañ Rosique, Profesor Titular de Cirugía de la Universidad Miguel Hernández y D. Carlos Alberto Mestres Lucio, Cardiothoracic and Vascular Surgeon, Department of Cardiothoracic and Vascular Surgery, Heart and Vascular Institute Cleveland Clinic, Abu Dhabi

CERTIFICAN:

Que la tesis titulada " VALIDACIÓN DE EuroSCORE II EN UN CENTRO DE MEDIO VOLUMEN" de la que es autor D. Antonio García Valentín, ha sido realizada bajo nuestra dirección.

Y tras valorar el trabajo realizado por el aspirante al Título de Doctor,

AUTORIZAMOS

Su presentación y defensa ante el Tribunal correspondiente.

Y para que conste a los efectos oportunos, firmamos el presente certificado en Alicante a 10 de Diciembre de 2015.

Fdo: Dr. D. Antonio F. Compañ Rosique

Fdo: Dr. D. Carlos Alberto Mestres Lucio

UNIVERSITAS
Miguel
Hernández

A mi mujer Alicia, por tu perenne sonrisa y tu apoyo incondicional. Por estar siempre.

A mis pequeños Pablo y María. Vosotros sois lo mejor de mi vida.

A mis padres Antonio y Copelia, por todo vuestro esfuerzo y sacrificio. Os lo debo todo.

A mi hermano Pablo, por ser mucho más que un hermano.

A mis suegros Josefina y especialmente a Pepe, por quererme y apoyarme como a su propia hija. Pepe, ací tenim per fi la tesi!

UNIVERSITAS
Miguel
Hernández

AGRADECIMIENTOS

A mis directores de tesis, los doctores Antonio Compañ y Carlos Mestres, los referentes más importantes en mis diferentes periodos de formación, con los que he tenido el honor y el placer de compartir la realización de este trabajo. Gracias por todo.

A mis antiguos compañeros del Hospital General Universitario de Alicante, cuyo trabajo ha servido para la realización de esta tesis; los cirujanos Patricio Llamas, Agustín Casillas, Aquilino Hurlé, Juan Meseguer, Jack Ventura y muy especialmente a mi amigo Eduardo Bernabeu, compañero de fatigas y proyectos desde hace muchos años. Al equipo de perfusionistas, Ana Carral, Emilio Sánchez y José Torres, por su gran esfuerzo y su inestimable ayuda en este trabajo. También a los cirujanos en formación que colaboraron en este proyecto, Rebeca Manrique y Yakir Castillo. Por último a todo el equipo de enfermería del Servicio, resto de especialistas y personal cuya labor se ve aquí reflejada.

A las personas que, desde la Facultad de Medicina a mi etapa en el Hospital Clínic, inculcaron en mí la curiosidad y valentía necesarias para culminar esta tarea, especialmente a Fito Aracil.

A las todos aquellos que me han ayudado y aconsejado en el campo de los modelos predictivos y evaluación de la calidad asistencial en Cirugía Cardiovascular, especialmente a Miguel Josa, José María Cortina, Jacobo Silva y mi querido amigo Daniel Pereda.

A mis antiguos compañeros del Hospital Clínico Univeristario de Valencia, que compartieron conmigo la aventura de la escritura de esta tesis, y a mis actuales compañeros del Hospital Universitario del Vinalopó, Eduardo Tébar, José Albors y Jorge Alcocer, que me han aportado ilusión y ayuda en la fase final.

UNIVERSITAS
Miguel
Hernández

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

UNIVERSITAS
Miguel
Hernández

ÍNDICE DE TABLAS

UNIVERSITAS
Miguel
Hernández

ABREVIATURAS

BNP	*Brain Natriuretic Peptide.* Péptido cerebral natriurético.
CCS	*Canadian Cardiovascular Society.* Sociedad Cardiovascular Canadiense.
CASS	*Coronary Artery Surgery Study*. Estudio de la Cirugía de las Arterias Coronarias.
CrCl	Aclaramiento de creatinina.
χ^2	Chi cuadrado.
DE	Desviación estándar.
DMID	Diabetes *mellitus* dependiente de insulina.
EACTS	European Association for Cardio-Thoracic Surgery. Asociación Europea para la Cirugía Cardio-Torácica.
EEUU	Estados Unidos de América.
EuroSCORE	*European System for Cardiac Operative Risk Evaluation.* Sistema Europeo para la Evaluación del Riesgo Operatorio Cardiaco.
FEVI	Fracción de eyección del ventrículo izquierdo.
HGUA	Hospital General Universitario de Alicante
HTP	Hipertensión pulmonar.
IAM	Infarto agudo de miocardio
MAR	Mortalidad ajustada al riesgo.
NY	Nueva York.
NYHA	*New York Heart Association.* Asociación del Corazón de Nueva York.
PARTNER	*Placement of Aortic Transcatheter Valves*. Colocación de Válvulas Aórticas Transcatéter.

ROC	Receiver Operating Characteristic. Característica Operativa del Receptor.
SECTCV	Sociedad Española de Cirugía Torácica-Cardiovascular.
STS	*Society of Thoracic Surgeons.* Sociedad de Cirujanos Torácicos.
USA	*United States of America*. Estados Unidos de América.
WA	Washington.

DEFINICIONES

Alto riesgo quirúrgico Paciente cuyo riesgo de mortalidad peroperatoria excede un porcentaje variable, entre un 7%-20% según los estudios, o que se encuentran en un tercil superior de la distribución de riesgo (mayor de seis en el EuroSCORE aditivo original).

Análisis univariante Prueba estadística que calcula el grado de asociación estadística de una variable con un evento de forma individual.

Análisis multivariante Prueba estadística que calcula el peso de cada variable dentro de un modelo mediante técnicas de regresión estadística.

Chi cuadrado Prueba estadística para contraste de hipótesis entre variables cualitativas.

Cirugía cardiaca mayor Intervención quirúrgica realizada para el tratamiento de patología estructural del corazón y/o grandes vasos torácicos que precisa apertura de la caja torácica.

Cirugía cardiaca menor Intervención quirúrgica realizada para el tratamiento de patología estructural del corazón y/o grandes vasos torácicos, así como otras técnicas sobre pared torácica, pericardio o dispositivos cardiacos, que no precisan apertura de la caja torácica.

Error α Probabilidad de rechazar la hipótesis nula cuando es cierta, o lo que el lo mismo, de obtener diferencias cuando realmente no existen.

Error β Probabilidad de rechazar la hipótesis alternativa cuando es cierta, o lo que el lo mismo, de no obtener diferencias cuando realmente existen. Inversa de la potencia estadística.

Endocarditis activa Variable discreta dicotómica recogida en EuroSCORE y EuroSCORE II, definida como paciente intervenido por endocarditis aguda mientras todavía está bajo tratamiento antibiótico por esa causa.

Estadístico c Valor del área bajo la curva ROC.

Estado crítico preoperatorio	Variable discreta dicotómica recogida en EuroSCORE y EuroSCORE II, definida como paciente intervenido en situación de ventilación mecánica, fallo renal agudo, resucitación reciente, arritmias malignas, masaje cardiaco, uso de fármacos vasoactivos o balón de contrapulsación intraaórtico.
Logit	Logaritmo de la *odds*.
Medicare	Sistema de cobertura de seguridad social financiado por el gobierno estadounidense
Método *forward*	Técnica de regresión en la que se genera un modelo inicial que se completa mediante la adición progresiva de nuevas variables.
Odds	Anglicismo empleado en estadística, infrecuentemente traducido al español como momio. Cociente entre número de opciones a favor y número de opciones en contra de un evento aleatorio, con significado similar, aunque no igual, al de probabilidad.
Odds ratio	Razón de oportunidades. Medida de asociación estadística que se emplea característicamente en estudios epidemiológicos de casos y controles.
Potencia estadística	Probabilidad de aceptar la hipótesis alternativa cuando ésta es cierta, o lo que es lo mismo, de obtener diferencias cuando realmente existen.
Razón O/E	Cociente entre la mortalidad observada y la esperada.
Sobreajuste	Defecto metodológico de un modelo predictivo, por el que se produce un ajuste excesivo a las características particulares de la población de desarrollo, presentando mal rendimiento en otras poblaciones.
Tamaño de la muestra	Número de observaciones que componen la muestra de una investigación. En nuestro medio, número de sujetos extraidos de una población que son incluidos en un estudio.
Válvulas transcatéter	Prótesis valvulares que se implantan mediante dispositivos intravasculares, por vía percutánea o mínimo acceso quirúrgico, sin necesidad de circulación extracorpórea.

Volumen quirúrgico Número de intervenciones realizadas por un centro, servicio o cirujano por unidad de tiempo. De acuerdo con datos extraídos de los registros de actividad más recientes, hemos considerado como centros de medio volumen en España a aquellos que realizan entre 250 y 400 intervenciones de cirugía cardiaca mayor anuales, considerando de alto volumen los que presentan una actividad mayor, y de bajo volumen los que presentan una actividad menor.

UNIVERSITAS
Miguel
Hernández

1. INTRODUCCIÓN

1.1. Diseño y elaboración de modelos predictivos

Los modelos clínicos de predicción pueden ser clasificados en dos categorías, los predictivos y los diagnósticos. Ambos comparten la característica de calcular la probabilidad de aparición de un suceso, diferenciándose principalmente en el aspecto temporal; los primeros son aquellos que utilizan características basales de los pacientes (factores de riesgo demográficos, datos clínicos o de la intervención prevista) para calcular la probabilidad de aparición de eventos futuros (mortalidad, complicaciones u otros). Su diseño es longitudinal en el tiempo, y entre ellos se incluyen las escalas de predicción de riesgo quirúrgico que empleamos habitualmente. Los segundos suelen tener un diseño transversal y tratan de calcular la probabilidad de emitir un diagnóstico concreto en base a unos datos clínicos o exploratorios, utilizando otra prueba diagnóstica altamente específica (patrón oro) para corroborar o no la presencia de la enfermedad (1).

Todos los modelos pueden ser construidos mediante diseños retrospectivos, que resultan más sencillos, pero con riesgo de pérdida de datos y de sesgo de selección, o prospectivos, que son más rigurosos en cuanto a los datos iniciales de cada paciente (2). Es preferible la recogida directa de datos de fuentes médicas frente a la utilización de registros administrativos no médicos, dado su bajo nivel de exhaustividad (3).

1.1.1. Selección de predictores y resultado

Las variables predictoras son aquellas que tienen una relación estadística con la aparición del evento resultado. Poseen un mayor o menor peso dentro del modelo dependiendo del grado de asociación estadística que tengan con el resultado a predecir, expresado en la forma correspondiente según el tipo de estudio epidemiológico, siendo el método más empleado la razón de oportunidades (también conocido por su término en inglés, *odds ratio*). Cualquier característica

demográfica del paciente, estadio funcional de la enfermedad, o característica de la intervención a realizar que presente una asociación estadística con el resultado puede ser un predictor.

El parámetro más importante para la obtención de predictores adecuados es la *fuerza* de los mismos. La fuerza es una función estadística que combina el grado de asociación entre un predictor y el evento resultado y la prevalencia del mismo dentro de la población estudiada. Es posible que un predictor sea más relevante (más fuerte) que otro teniendo una asociación estadística más débil con el evento final, por ser mucho más prevalente, y al contrario. Un ejemplo es la repetida ausencia de un predictor del tipo "Disfunción Hepática" en los modelos de predicción de riesgo quirúrgico más empleados. A pesar de que la presencia de una cirrosis hepática en estadio avanzado está asociada a una alta mortalidad peroperatoria (4-7), su escasa prevalencia entre los pacientes habitualmente sometidos a cirugía cardiaca ha eliminado esta variable de las ecuaciones finales de EuroSCORE (*European System for Cardiac Operative Risk Evaluation*) o *STS (Society of Thoracic Surgeons) Score* (8, 9). Otro ejemplo es la desaparición de la variable "Rotura del Septo Interventricular Post-Infarto" como predictor aislado en EuroSCORE II por el mismo motivo (10). Cada predictor ha de ser sometido inicialmente a un análisis univariante en busca del grado de asociación individual con el resultado a predecir, aplicando los métodos estadísticos adecuados para esta tarea en función del diseño del estudio, y que precisará de un tamaño de la muestra adecuado en función de los errores α y β, de la prevalencia del predictor y de la magnitud del efecto a observar. Es importante definir con rigor cada variable de predicción para evitar la variabilidad entre observadores y el consiguiente sesgo de dilución a la hora de generar el modelo predictivo.

Cualquier evento beneficioso o adverso relacionado con la intervención quirúrgica puede ser el resultado cuya probabilidad de aparición se quiere calcular. Es preferible elegir resultados bien definidos y de gran trascendencia clínica, llamados resultados *duros* (2). El más empleado en Cirugía Cardiovascular es la mortalidad, aunque algunas complicaciones no letales pueden ser importantes

también por su gran impacto en la calidad de vida del paciente o por motivos económicos (accidente cerebrovascular, mediastinitis, diálisis postoperatoria, etc.) También es posible emplear resultados compuestos que combinen mortalidad con eventos no fatales, como ocurre en el *STS Score*. En la mayoría de modelos de predicción empleados en Cirugía Cardiovascular se ha empleado como variable resultado la mortalidad peroperatoria, definida como aquella ocurrida durante el ingreso hospitalario (con independencia de su duración) o hasta 30 días tras la intervención. Sin embargo en EuroSCORE II este objetivo ha sido redefinido como mortalidad exclusivamente intrahospitalaria por la importante pérdida de datos en el seguimiento de los pacientes del estudio (10), decisión que no ha sido ajena a la polémica (11).

Un punto de importancia en la elaboración de modelos predictivos es el manejo de los datos perdidos en la muestra utilizada para su desarrollo. La pérdida de datos puede afectar tanto a las variables predictoras como a la variable resultado. Cuando ocurre entre las variables predictoras la consecuencia es la ineficiencia del trabajo de investigación, dado que la pérdida de pacientes provoca una disminución del tamaño de la muestra, generando un sesgo de dilución en los coeficientes de las variables. Existen métodos estadísticos para completar los datos perdidos, llamados métodos de imputación, que se basan en la relación directa de unos predictores con otros (por ejemplo, un valor de hemoglobina alto está correlacionado con un valor de hematocrito alto), u otros métodos que asumen una distribución concreta de los datos dentro de la muestra. Sin embargo la pérdida de datos relacionada con la variable resultado es mucho más grave, ya que nos impide calcular la asociación entre predictores y resultado, conduciendo al cálculo de coeficientes erróneos y modelos no válidos. También es posible realizar técnicas de imputación, aunque no resultan útiles en estos casos porque aplican las relaciones predictor-resultado calculadas con los datos existentes.

1.1.2. Bases estadísticas de la construcción de modelos predictivos

El desarrollo estadístico necesario para elaborar un modelo predictivo varía según el tipo de resultado buscado. Para resultados definidos por variables continuas se emplea el modelo de mínimos cuadrados o regresión lineal. Este modelo se construye de forma que el resultado *y* es una combinación lineal de *n* variables *x*, según la expresión $y=\alpha+\beta_n x_n$, donde α es el intercepto u ordenada en origen (intersección de la recta obtenida con el eje de ordenadas) y β_n son los coeficientes multiplicadores de cada variable predictora x_n. A mayor coeficiente, mayor asociación estadística entre predictor y resultado. La representación gráfica de esta regresión puede apreciarse en la Figura 1.

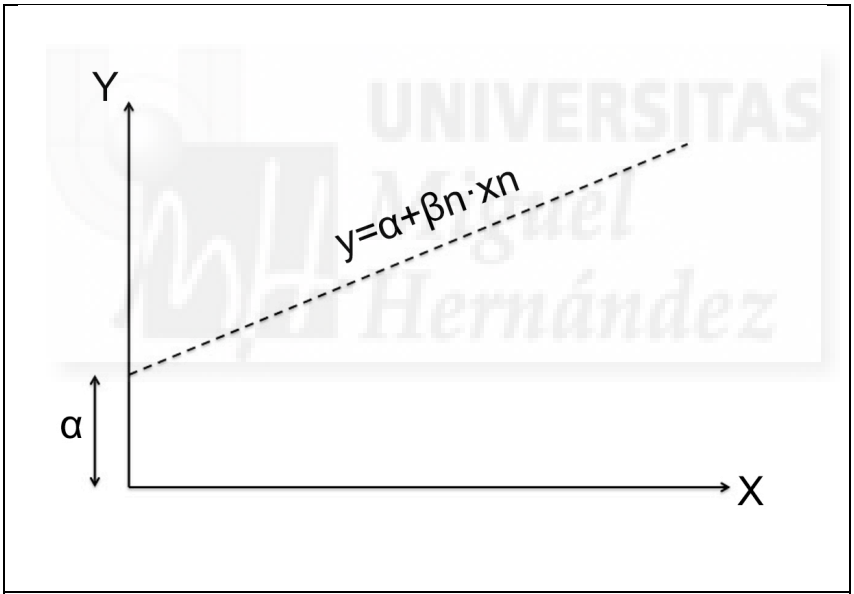


Figura 1. Representación gráfica de la regresión lineal.

En los modelos de predicción del riesgo quirúrgico la variable resultado más empleada es la mortalidad, que es de tipo binario (discontinua dicotómica). El modelo estadístico empleado para la construcción de modelos con este tipo de variable resultado es la regresión logística, que emplea la función *Logit* (logaritmo de la *odds*). Se construye expresando el *Logit* de la probabilidad de que ocurra un evento *y* (P(y=1)) como el resultado de la combinación lineal de los predictores. Dado que la *odds* es igual a la probabilidad de que ocurra el suceso dividida entre

la probabilidad de que no ocurra (la *odds* de obtener un 6 en una tirada de dados sería de 1/5), podemos formular la expresión descrita de la siguiente manera:

$$\textit{Logit}\ (P(y=1)) = \log \textit{odds}\ (P(y=1)) = \log \frac{P(y=1)}{1-(P(y=1)} = \alpha+\beta_n x_n.$$

La probabilidad de aparición del evento *y* puede ser despejada, por tanto, mediante la siguiente expresión:

$$P(y=1) = \frac{e^{\alpha+\beta n \cdot xn}}{1+e^{\alpha+\beta n \cdot xn}} = \frac{1}{1+e^{-(\alpha+\beta n \cdot xn)}}$$

Esta transformación nos permite restringir los valores de la función al intervalo (0, 1), siendo su representación gráfica la curva logística, con una forma sigmoidea característica (Figura 2). El coeficiente β_n representa la medida de la asociación entre cada uno de los predictores y el evento resultado, de forma que la *odds ratio* de cada predictor es en número *e* elevado a su correspondiente coeficiente β (*odds ratio* $x_n = e^{\beta n}$).

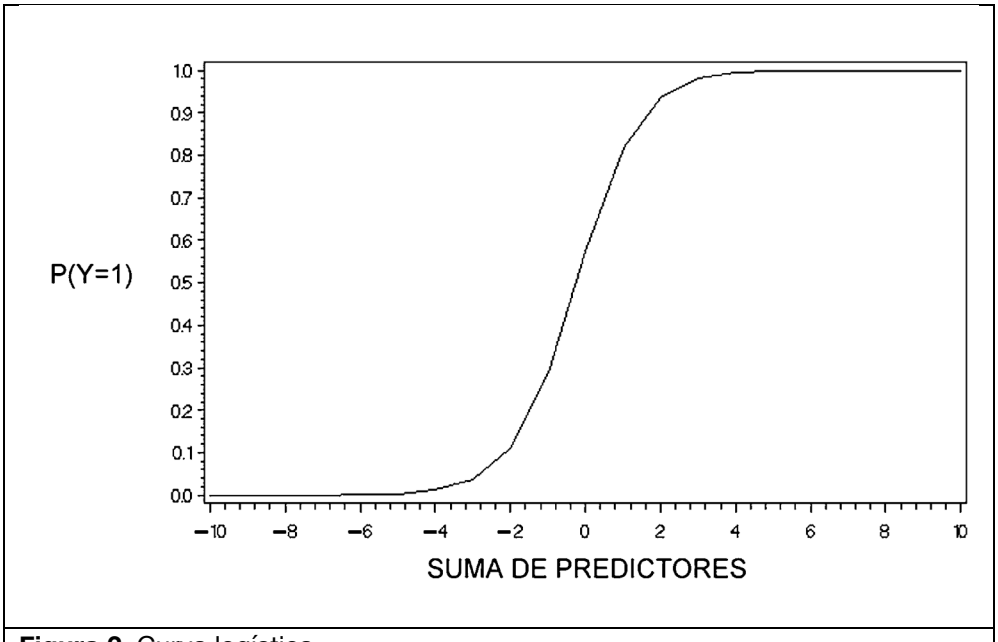


Figura 2. Curva logística.

Una variante más compleja de las regresiones logísticas son las llamadas redes neurales, ya empleadas en robótica y computación, que tienen capacidad de aprendizaje y reajuste de los coeficientes mediante mecanismos de retroalimentación, y que probablemente sean la base de futuros modelos de predicción en el campo de la Cirugía Cardiovascular (10).

Existe otro tipo de resultados, muy frecuentes en la literatura, que son aquellos dependientes de tiempo. Ejemplos de ello son los resultados de supervivencia de una muestra de pacientes sometidos a una técnica quirúrgica, o las series de supervivencia libre de reintervención en pacientes a los que se realiza una reparación valvular. Dado que los tiempos de seguimiento son variables y existen datos censurados (incompletos), correspondientes a pérdidas en el seguimiento, es necesario un método estadístico específico para su predicción, siendo el modelo de riesgo proporcional de Cox el más empleado. Ésta no es una función lineal, y calcula el exceso de riesgo de aparición del evento resultado a lo largo del seguimiento, en referencia a una función de riesgo basal. El análisis de Kaplan-Meier también se emplea con mucha frecuencia en análisis de supervivencia, y su estadístico *log-rank* resulta útil para mostrar relaciones univariantes de los predictores con el evento.

Las variables predictoras pueden ser continuas o categóricas (pudiendo a su vez ser estas últimas dicotómicas, si pueden tomar solamente dos valores, o politómicas, si pueden tomar más de dos). Las dicotómicas son las más sencillas de interpretar, ya que si se encuentran presentes toman un valor 1 que se multiplica por su correspondiente coeficiente, mientras que si no existen toman un valor 0. Las variables politómicas o multicategóricas precisan del estudio de la asociación de cada una de los valores que puede tomar la variable, respecto al valor menos asociado al resultado, que será la categoría de referencia. La categoría de referencia puede ser la de mayor o menor valor, dependiendo de las implicaciones clínicas o fisiológicas de cada variable (la presión pulmonar baja se se considera normal y la alta patológica, mientras que la interpretación del valor de la fracción de eyección ventricular es completamente opuesta.) Se obtienen así

coeficientes diferentes para cada una de las categorías, que se comportan como predictores independientes, aunque no es posible su aparición simultánea. EuroSCORE II contiene 6 predictores multicategóricos construidos de este modo. Tomando como ejemplo la variable "Fracción de Eyección" (FE), apreciamos 4 categorías: FE>50%, que es la categoría referencia y no tiene asignado coeficiente, por lo que no incrementa la probabilidad de mortalidad; FE 30-50%, que tiene un coeficiente β de 0,315 con respecto a la anterior; FE 20-30%, con un coeficiente β de 0,808; y FE<20%, que tienen un coeficiente β de 0,934 con respecto a la referencia. Por último tenemos las variables continuas, siendo la edad la más habitual en los modelos de predicción de riesgo quirúrgico. Normalmente se define una edad mínima que no condiciona incremento de riesgo para valores inferiores. A partir de ese punto se calcula un coeficiente que se multiplica por el número de años que exceden del valor base (edad mínima), incrementando el riesgo de mortalidad.

1.1.3. Problemas de diseño de los modelos predictivos

Existen problemas intrínsecos a la metodología empleada en la elaboración de modelos predictivos, cuya importancia reside en que pueden comprometer la validez de sus resultados. Como usuarios habituales de estas herramientas debemos conocer, comprender e identificar estos problemas a la hora de evaluar el rendimiento de estos modelos en nuestros pacientes o en estudios de validación.

El sobreajuste (*overfitting*) ocurre por un exceso de grados de libertad en la elaboración del modelo, por ejemplo, por la existencia de demasiadas variables predictoras, o por el desarrollo del mismo a partir de muestras de pequeño tamaño (12). Esto produce que el modelo se ajuste en exceso a las características particulares de la muestra empleada para su elaboración, alejándose de ese modo de las características de la población general, donde tendrá peor rendimiento. El concepto derivado de este problema se denomina optimismo. Éste se define como la diferencia entre el rendimiento verdadero de un modelo en

la población real y el rendimiento aparente en la muestra empleada para su elaboración. Otra de las causas principales de sobreajuste y optimismo es la incertidumbre, que en el campo de la predicción se define como la ausencia de un conjunto de variables predefinidas antes de desarrollar el modelo. La inclusión de un gran número predictores candidatos hace necesarios múltiples análisis para hallar asociaciones estadísticas adecuadas. Esto puede ocasionar que se obtengan variables supuestamente asociadas por puro azar, incluyendo así en el modelo final predictores que son solamente ruido. El optimismo puede disminuir el rendimiento de una escala, y puede ser reducido con tamaños de muestra muy grandes. Una forma de reducir el número de variables candidatas a predictor se basa en el conocimiento propio y en la consulta previa de la literatura, que nos puede orientar hacia la elección de variables cuya asociación ha sido descrita ya anteriormente. Aun así en la práctica resulta imposible conocer *a priori* si todas las variables tendrán o no una asociación real con el evento resultado, por lo que a menudo es necesario efectuar múltiples análisis, con los riesgos de sobreajuste y optimismo ya descritos.

Existen diversos métodos que permiten detectar y calcular el grado de optimismo. Probablemente el más aplicado es *bootstrap*, cuya traducción literal es “correa de bota”, nombre derivado de uno de los episodios de las aventuras del Barón de Münchhausen. En él, el protagonista lograba escapar de un pantano estirando de la correa de sus propias botas (13), lo que refleja de forma intuitiva la filosofía de este método estadístico, que utiliza la propia población de desarrollo del modelo para analizar su rendimiento. *Bootstrap* es un método que utiliza toda la población de desarrollo, seleccionando de entre ella múltiples muestras aleatorias para reconstruir el modelo varias veces con los pacientes de cada selección. De esta forma se pueden construir múltiples modelos con diferentes coeficientes β, y evaluar el rendimiento de cada nuevo modelo con respecto al modelo original. La diferencia media de rendimiento entre estos modelos y el original será el optimismo (12).

Además de los problemas inherentes al proceso matemático empleado para el desarrollo de modelos predictivos, existen numerosos sesgos asociados a cualquier estudio científico que pueden sumarse a estos, limitando el rendimiento del mismo. El sesgo de selección es aquel en el que la selección de los predictores se ve afectada por la disponibilidad de los datos, y es más acusado en estudios de diseño retrospectivo. El sesgo de observación (también llamado efecto Hawthorne) consiste en la alteración de los resultados de un grupo (productividad, mortalidad, etc.) al conocer que están siendo estudiados (14-16). El sesgo de publicación hace referencia a que los resultados publicados suelen ser mejores o más atractivos (mayores diferencias o tamaños de muestra) que aquellos no publicados, que por otro lado pueden ser más acordes a la realidad de la práctica habitual (17-20). Todos estos factores son importantes al ser capaces de limitar el rendimiento de un modelo, por lo que han de ser tenidos en cuenta.

El último concepto que habitualmente se menciona en las normas de elaboración de modelos predictivos es el principio de parsimonia, que defiende que la explicación completa más simple suele ser la mejor. Esto implica que modelos con menor número de predictores deben rendir adecuadamente. Aunque esto puede ser puesto en duda, sí es cierto que en la práctica las escalas más simples son más fáciles de recordar y emplear (2).

Como resumen, un modelo predictivo debe contener entre 5 y 20 predictores con fuerza y rigurosamente definidos para ser eficaz, controlando sesgos, confusores y efectos de interacción.

1.1.4. Rendimiento y validación de los modelos predictivos

El rendimiento de un modelo predictivo es un concepto que se refiere a la calidad de las predicciones que realiza. Puede ser valorado de forma global por diversos métodos estadísticos (R^2 de Nagelkerke, *score* de Brier, índice de reclasificación, etc.) (21), aunque generalmente diferenciamos dos aspectos dentro de este

objetivo: la calibración y la discriminación del modelo. Por calibración se define el grado de acuerdo entre los resultados esperados y observados; en concreto para modelos de predicción de riesgo quirúrgico, valorar si la mortalidad media pronosticada se ajusta a la que observamos en la población del estudio de validación. Por discriminación se entiende la habilidad del modelo para identificar a aquellos individuos que presentarán el evento resultado, en nuestro medio, la capacidad del modelo para identificar a los pacientes que no sobrevivirán a la intervención quirúrgica. Los métodos estadísticos más empleados en la literatura actual para valorar la calibración y la discriminación de un modelo son la prueba de bondad de ajuste de Hosmer-Lemeshow y el área bajo la curva ROC (*receiver operating characteristic*), respectivamente. Ante un modelo con buenos parámetros de rendimiento en una muestra poblacional, normalmente concluimos que está validado para dicha población.

Aunque debatida, la prueba de bondad de ajuste de Hosmer-Lemeshow (22) sigue siendo el estadístico más empleado para la evaluación de la calibración en los modelos de predicción de riesgo quirúrgico (23). Se trata de un estadístico de contraste basado en la prueba χ^2 (Chi cuadrado), en la que de forma habitual se agrupa a los pacientes por deciles de riesgo de aparición de la variable resultado (0%-10%, 10%-20%, etc.), aunque pueden utilizarse diferente número de divisiones. Se calcula para cada grupo la diferencia entre valores observados y esperados, y se agrupan los resultados en un solo valor de p final. Es un método comprensible e intuitivo, ya que implica que no sólo la mortalidad global observada debe ser similar a la esperada, sino que además debe serlo en cada uno de los grupos de riesgo. El valor de p que se emite como resultado indica la probabilidad de que las diferencias existentes entre eventos observados y esperados sean debidas solo al azar. Para concluir que un modelo predictivo tiene una buena calibración nos interesa, por tanto, que el valor de p obtenido por la prueba de Hosmer-Lemeshow sea lo más cercano a uno posible. A pesar de que como norma se acepta como límite de significación estadística el valor de p menor de 0,05, teóricamente no podemos concluir que una escala está correctamente calibrada simplemente por presentar un valor de p mayor de 0,05.

La interpretación más adecuada del valor de p en la prueba de Hosmer-Lemeshow es que el modelo predictivo estará mejor calibrado cuando mayor sea el valor de p. Como ya se mencionó, hoy en día la precisión de la prueba de bondad de ajuste de Hosmer-Lemeshow se encuentra bajo debate (24). Existen varios puntos en que esta prueba resulta controvertida desde el punto de vista estadístico; el uso sistemático en la literatura de deciles de riesgo, cuando se pueden utilizar otro tipo de divisiones como cuartiles, quintiles o percentiles, y la alta probabilidad de obtener fallos de calibración al emplear conjuntos de datos demasiado voluminosos, por la gran sensibilidad del χ^2 al tamaño de la muestra. Respecto a estos puntos, el trabajo que Paul firma justo a Lemeshow concluye que el número de grupos debe ser de 10 para tamaños de hasta 1.000 pacientes, debiendo ajustarse a otras cantidades para grandes conjuntos de datos (entre 1.000 y 25.000 pacientes). Esta premisa conlleva asimismo consideraciones particulares en el número de grupos cuando la probabilidad del evento es baja (menor de 0,1), circunstancia en la que la prueba de bondad de ajuste puede rendir mal. También se concluye en el mismo artículo que la prueba de bondad de ajuste de Hosmer-Lemeshow no debe ser utilizada para conjuntos de datos superiores a 25.000 pacientes, así como la escasa capacidad de esta prueba de bondad de ajuste para detectar problemas como el sobreajuste y el optimismo (25).

Existen otros métodos de calibración más perfeccionados (pendiente de calibración, *calibration-in-the-large*, etc.), cada vez más empleados en estudios de validación actualmente, aunque todavía no muy frecuentes.

La curva ROC es una representación gráfica de la sensibilidad de un modelo predictivo (verdaderos positivos, eje de ordenadas) frente a 1-especificidad (falsos positivos, eje de abscisas) para cada uno de los pacientes. Estas curvas se comenzaron a emplear para el análisis de señales de radar durante la segunda guerra mundial, con el objetivo de averiguar el poder discriminativo de las señales para la detección de objetivos, dando lugar a la llamada Teoría de la Detección de Señales. Actualmente esta prueba supone en Medicina la principal herramienta en

el análisis de discriminación de modelos predictivos (12) y puntos de corte en pruebas diagnósticas (26). El dato que nos informará sobre la habilidad discriminatoria del modelo es el área bajo la curva ROC (estadístico *c*). Un modelo predictivo perfecto estaría representado por una línea que delimitaría un cuadrado en el que todos los positivos serían verdaderos, con un área bajo la curva de 1. Un modelo sin capacidad real de predicción (por ejemplo, lanzar una moneda al aire) dibujaría una línea diagonal con un área bajo la curva de 0,5 (Figura 3).

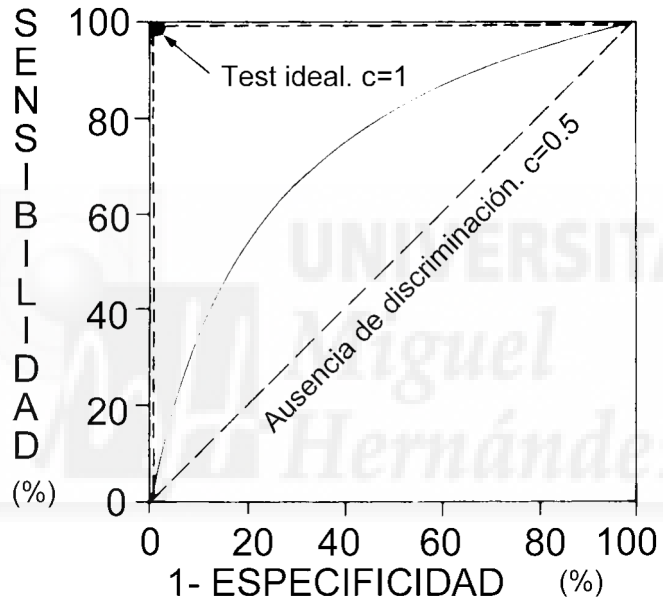


Figura 3. Curva ROC y líneas de referencia.

Por tanto, cuanto más se aproxime el valor de *c* a 1, mayor capacidad de discriminación tendrá. Se consideran adecuados los valores de *c* superiores a 0,7, siendo considerados buenos valores superiores a 0,8 y muy buenos los superiores a 0,9 (3). El valor de *c* debe acompañarse de un intervalo de confianza (normalmente al 95%), que será más estrecho y preciso cuanto mayor sea el tamaño de la población del estudio de validación, y por tanto menor la desviación estándar (27). Es fundamental que el intervalo de confianza no incluya valores cercanos a 0,5, lo que implicaría la ausencia total de discriminación del modelo.

La validación de modelos predictivos se basa en la evaluación del rendimiento en distintas poblaciones, y nos da una idea acerca de la capacidad de generalización de dicho modelo (21). En primer lugar es necesario comprobar la reproducibilidad del modelo evaluando la validez interna del mismo, es decir, comprobando su rendimiento sobre la misma población que se utilizó para generarlo. Posteriormente se procede a evaluar el rendimiento del modelo en distintas poblaciones con diferentes características a las de la población generadora en un proceso de validación externa, que nos dará la característica de modelo generalizable. En estudios de validación, la hipótesis nula (H_0) concluye que no hay diferencias entre datos observados y esperados, y el modelo predictivo funciona correctamente.

Para realizar la validación interna de un modelo se emplea la misma población de sujetos que se utilizó para su desarrollo. Puede utilizarse la población entera, o dividirla aleatoriamente en un grupo de desarrollo del modelo y otro de validación, como se realizó en los trabajos originales de EuroSCORE y EuroSCORE II (8, 10), aunque esta técnica puede generar desequilibrios importantes entre grupos. Otro método es escoger múltiples subpoblaciones aleatorias de la muestra para recrear sucesivos modelos con sucesivas validaciones internas, en un modelo de validación cruzada (empleada también en EuroSCORE II para valoración de la estabilidad) (10). La técnica *bootstrap,* ya comentada anteriormente también se puede emplear para la evaluación de la validación interna, generando subgrupos aleatorios cuyo rendimiento medio será el calculado para la validación interna.

La validación externa implica la aplicación de un modelo a una población distinta de la que se utilizó para desarrollarlo. Estas diferencias pueden tener un origen temporal (pacientes intervenidos en otro momento de la evolución de la enfermedad), geográfico (validación en otro centro, ciudad o país), etc.

Un estudio de validación precisa un cálculo óptimo del tamaño de la muestra. Para el cálculo del área bajo la curva ROC necesitamos un número mínimo de eventos para obtener un intervalo de confianza del estadístico *c* lo más estrecho

posible a fin de alejarse del valor 0,5. Los únicos factores que influyen en este cálculo son el número de pacientes incluidos y el valor del área bajo la curva ROC estimada, como se demuestra en uno de los principales trabajos de referencia de esta técnica estadística (27). Por ejemplo, si calculamos que necesitamos 100 eventos (fallecimientos postoperatorios) para valorar la discriminación de un modelo de predicción de mortalidad en cirugía coronaria, y asumimos que la mortalidad media observada será aproximadamente del 2,5%, necesitaremos incluir 4.000 pacientes en nuestro estudio de validación. Los estudios diseñados para la comparación de dos áreas bajo la curva ROC para dos pruebas diagnósticas diferentes requieren asimismo tamaños de muestra mucho más grandes, de incluso miles o decenas de miles de pacientes, según el grado de diferencia entre los valores de *c* que quiera ser observado (28).

El cálculo del tamaño de la muestra para la prueba de bondad de ajuste, como ya se mencionó previamente, contiene problemas inherentes a su diseño, ya que veces no es posible distribuir la población del estudio de validación en deciles de riesgo homogéneos (en una validación de una escala de mortalidad en cirugía cardiaca la inmensa mayoría de los pacientes tendría valores de mortalidad esperada entre 0% y 10%). Por tanto, son precisos tamaños de muestra que pueden alcanzar cientos o miles de pacientes para obtener un número de eventos y una potencia estadística adecuada, con las salvedades antes descritas en el número de pacientes y grados de libertad. Como se puede deducir, es especialmente importante valorar los tamaños de la muestra e intervalos de confianza para la lectura crítica de los trabajos de validación de modelos de predicción, ya que muchas veces se realizan con pequeñas muestras de pacientes y escaso número de eventos, y por tanto ofrecen atractivos valores centrales con una amplísima dispersión, que los hacen ineficaces a la hora de extraer conclusiones (29, 30).

Las causas de no validación de un modelo predictivo son múltiples. Un estudio con un área bajo la curva ROC o una prueba de Hosmer-Lemeshow que indican un rendimiento bajo normalmente concluye que un modelo predictivo no tiene

validez para la población en la que se aplica, o que está mal diseñado. Sin embargo debemos valorar otras causas diferentes de no validación, por ejemplo un mal diseño del estudio de validación por insuficiente tamaño de la muestra, o por la existencia de sesgos de selección o información por diseño retrospectivo del mismo. Asimismo una causa importante y siempre difícil de admitir, en los casos de fallo de calibración por infraestimación del riesgo de mortalidad, es que en efecto existan diferencias reales entre el rendimiento predicho por el modelo y el del centro, región o país donde se quiere validar. El rendimiento de un equipo quirúrgico es un concepto complejo y sometido a múltiples variables, algunas de ellas poco cuantificables, como el buen funcionamiento de los equipos de Cardiología, la optimización de las indicaciones quirúrgicas, la adecuada provisión por parte de la administración de materiales y equipos adecuados para los cuidados intra y postoperatorios, etc.

1.2. Mortalidad ajustada al riesgo y gráficas de embudo. Uso de los modelos predictivos en la evaluación de la calidad asistencial

La mortalidad ajustada al riesgo es la herramienta que nos permite evaluar el rendimiento quirúrgico de una unidad asistencial, utilizando como parámetro de calidad la mortalidad peroperatoria. Se basa en el cálculo de la razón O/E, que es el cociente de la mortalidad observada (O) entre la esperada (E). Ésta última se calcula a través de la medida de tendencia central correspondiente (media o mediana, según la distribución de la muestra) de las mortalidades esperadas calculadas para todos los pacientes estudiados, empleando para ello un modelo predictivo validado en esa población. Una razón O/E igual a 1 implica que la mortalidad observada es igual a la esperada, y que el rendimiento del equipo quirúrgico es igual al esperado por una escala de riesgo seleccionada. Una mortalidad inferior a 1 implicaría que la mortalidad observada sería menor a la esperada, y que el funcionamiento del servicio sería mejor que el pronosticado por el modelo predictivo. Una mortalidad mayor de 1 implicaría deficiencias en el rendimiento del servicio, con mortalidades peroperatorias mayores de las esperadas por el modelo. Valores excesivamente lejanos a 1 deben hacer

sospechar de problemas metodológicos en los cálculos, antes que en un excelente rendimiento del Servicio, cuando la mortalidad ajustada al riesgo es excesivamente baja, o sugerir un problema de rendimiento de un servicio, si es muy alta (2).

Para estudios de comparación de los resultados de un grupo con una mortalidad de referencia (por ejemplo, de registros nacionales o internacionales previos), habitualmente se aplica la metodología propuesta por Spiegelhalter (31). El cálculo de la mortalidad ajustada al riesgo de una unidad quirúrgica se calcula como el producto de la mortalidad cruda de referencia (nacional, internacional) por el cociente O/E calculado para ese servicio en la muestra de estudio. Por ejemplo, si queremos ajustar al riesgo la mortalidad de un servicio quirúrgico con una mortalidad observada del 5%, siendo la mortalidad media esperada del 6%, el cociente O/E será de 5/6=0,83. Si la mortalidad cruda española es del 3%, la mortalidad ajustada al riesgo del servicio analizado será 3%·0,83=2,5%. Este cálculo asume que el modelo predictivo empleado está adecuadamente validado, y que por tanto la mortalidad observada y esperada deben ser similares en la población de referencia, siendo el cociente O/E de la misma igual a 1. Este cálculo simplificado puede dar lugar a situaciones matemáticamente aberrantes, por ejemplo en patologías con alta mortalidad peroperatoria, como por ejemplo la reparación de las roturas septales postinfarto. Supongamos una mortalidad de referencia del 40% en la población española o europea. Si la mortalidad esperada media fuese del 20% según EuroSCORE, y la observada fuese del 60% (en una patología tan infrecuente y con tan alta letalidad, el fallecimiento de un solo paciente puede incrementar de forma muy importante la mortalidad observada), el cociente O/E sería de 3, que multiplicado por la mortalidad de referencia nos daría una mortalidad ajustada al riesgo del 120%, hecho que es imposible. Para evitar estas situaciones, existe un ajuste en base a la *odds* de la mortalidad de referencia (comunicada al autor por Spiegelhalter) que limita los valores de la mortalidad ajustada al riesgo al rango 0-100%, evitando así la existencia de probabilidades de mortalidad imposibles desde el punto de vista empírico.

Las gráficas de embudo son una representación gráfica de estos cálculos, propuestas por el mismo autor como un buen método para la evaluación de la calidad asistencial en servicios sanitarios (31). Su existencia no es reciente, y ya han sido empleadas previamente en diversos estudios de control de calidad (32-36).

Estas gráficas se construyen como una representación de la mortalidad ajustada al riesgo en el eje de ordenadas, frente a un criterio de precisión en el eje de abscisas, que en nuestro campo es el número de pacientes intervenidos. Se representan habitualmente dos valores de mortalidad; la de referencia, que como ya se comentó, se extrae de la literatura a partir de trabajos y registros previos, se representa como una línea paralela al eje de abscisas, calculándose los límites del límite de confianza y alarma de forma simétrica a su alrededor; la mortalidad ajustada al riesgo del grupo estudiado se representa como un punto cuyo valor en el eje de abscisas es el número de pacientes incluidos en el estudio, y su valor en el eje y es el de la mortalidad ajustada calculada de la forma que se describió previamente. La gráfica con la característica forma de embudo (Figura 4) se construye a partir del cálculo de los límites de confianza, que se representan gráficamente como dos líneas simétricas especulares al eje de la línea de la mortalidad de referencia. Inicialmente los puntos que la componen están más alejados de ésta, y según aumenta el tamaño de la muestra y avanzamos por el eje de abscisas, se van acercando hasta crear un intervalo estrecho y asintótico a la línea de referencia. El hecho de que los intervalos de confianza se estrechen al mismo tiempo que aumenta el tamaño de la muestra se basa en que éste es el parámetro de precisión es que se basa el cálculo de dichos intervalos.

Es posible utilizar varios tipos de intervalo de confianza para el cálculo de estas gráficas, de la misma manera en que se emplean para la comunicación de resultados en estadística descriptiva o en medidas de asociación estadística. Es posible construir estos intervalos con una desviación estándar (al 68%), dos desviaciones estándar (aproximadamente al 95%) o 2,5 desviaciones estándar (aproximadamente al 99%) desde la mortalidad de referencia, siendo este último

el más utilizado en la literatura (31). El intervalo de confianza superior en el eje de ordenadas indica el límite de alarma por mortalidad excesiva, de forma que si el punto que representa la mortalidad ajustada al riesgo está por encima de este límite hemos de sospechar un bajo rendimiento quirúrgico de la unidad estudiada. Por el contrario un valor de mortalidad ajustada al riesgo por debajo del límite inferior de alarma indica, en condiciones normales, un excelente rendimiento quirúrgico del centro estudiado. Como se ha comentado previamente, es importante saber que los valores extremos pueden ser debidos a otras circunstancias, como deficiencias en los modelos predictivos (modelos que causen sobreestimación originarán mortalidades ajustadas al riesgo muy bajas, muy por debajo del los límites de alarma inferiores, mientras que modelos que generen infraestimación generarán lo contrario), o alteraciones en el cálculo del riesgo de los pacientes (como sobreestimar de forma intencionada las probabilidades de fallecimiento de los pacientes) (37). Por ello, en teoría lo recomendable sería comprobar el funcionamiento real de aquellas unidades cuya mortalidad ajustada quede por fuera de los intervalos de confianza, ya sea el superior o el inferior.

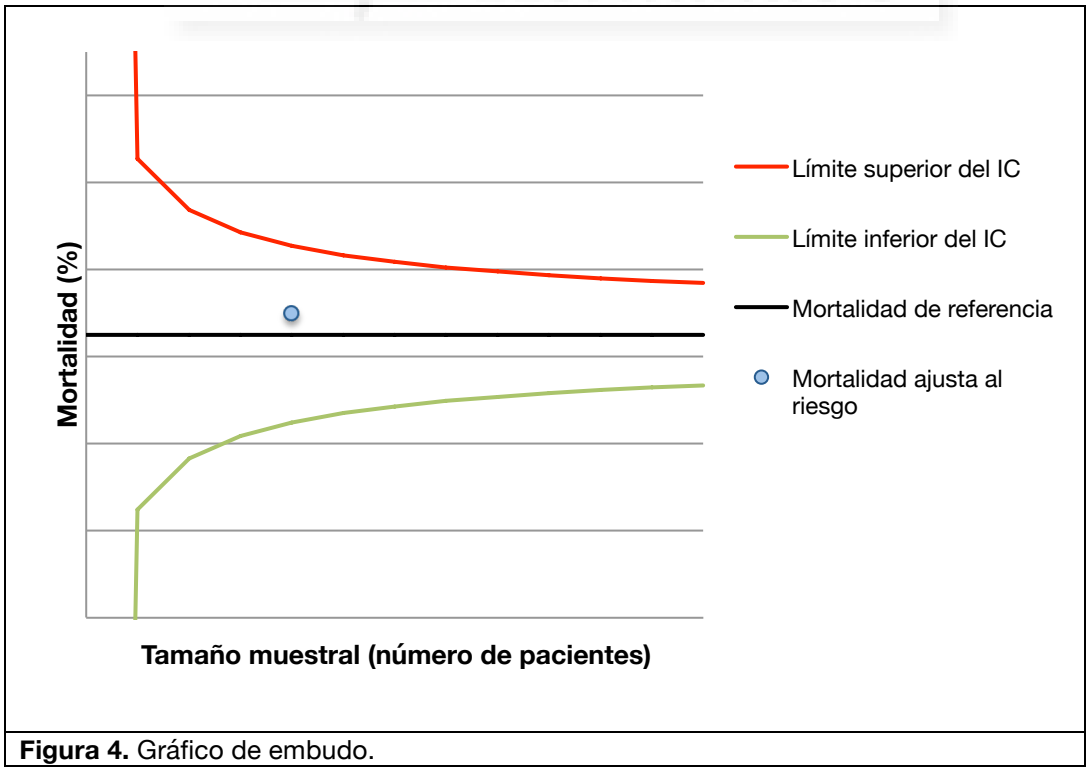


Figura 4. Gráfico de embudo.

1.3. Modelos predictivos en Cirugía Cardiovascular

La Cirugía Cardiovascular es una especialidad médica que ha demostrado máxima inquietud a lo largo de su historia por la recogida y análisis de los datos derivados de su actividad. Fruto de esta labor son los múltiples estudios existentes en las distintas publicaciones especializadas sobre aspectos epidemiológicos, resultados, técnicas quirúrgicas, etc.

Debido a las características intrínsecas de esta cirugía, en la que existe un riesgo no despreciable de mortalidad asociada a la intervención, y a la inquietud por conocer cuándo la indicación de cirugía es adecuada y qué riesgos y beneficios podemos ofrecer a los pacientes, surge la necesidad de crear modelos de predicción con los que comparar los resultados observados en nuestra práctica.

A lo largo de la aún joven historia de la Cirugía Cardiovascular se han elaborado distintos modelos de predicción del riesgo quirúrgico. Probablemente el modelo pionero sea aquél enmarcado dentro del estudio CASS *(Coronary Artery Surgery Study)*, que asignaba un riesgo de mortalidad operatoria en cirugía coronaria en base a parámetros clínicos y angiográficos (38).

Probablemente el siguiente modelo en importancia de entre los pioneros fuera el de Parsonnet, que incluía en su diseño multitud de variables para garantizar el máximo ajuste a las características particulares de cada paciente de forma individualizada (39). Este modelo fue utilizado y validado, aunque cayó en desuso por ciertos problemas de definición de variables y por la aparición de nuevos modelos realizados sobre bases de datos más amplias (3). En Norteamérica se han desarrollado diversos modelos adaptados a la práctica de centros específicos, generalmente hospitales de alto volumen que disponen de una gran cantidad de datos propios que les permiten crear así sus propios modelos de predicción. A pesar de ello el modelo más utilizado en América del Norte es el de la *Society of Thoracic Surgeons* (STS).

La labor realizada por la STS con su base de datos y el desarrollo de su modelo predictivo se remonta al año 1989. Desde entonces se han recogido los datos de los pacientes intervenidos en cirugía cardiaca de adultos, infantil y torácica en una muestra importante de centros, dentro de una iniciativa de esta sociedad por la mejora de la calidad asistencial y la seguridad del pacientes. Desde el año 2011 el modo de comunicación de datos es en línea, a través de una página de internet. Actualmente existen registros de más de 4.500.000 pacientes procedentes de más de 250 centros norteamericanos, lo que se calcula que representa aproximadamente el 94% de la actividad quirúrgica del país. *STS Score* es revisado y ajustado periódicamente, habiéndose efectuado su última actualización en 2009. Tiene capacidad de predicción del riesgo de mortalidad y morbilidad mayor (insuficiencia renal aguda, accidente cerebrovascular, ventilación mecánica prolongada, infección profunda de la herida esternal y reoperación) durante el ingreso postoperatorio y a 30 días, así como un evento combinado de morbimortalidad (40). Las características de este modelo de predicción se han ido conformando a lo largo del tiempo, y se reajustan con el crecimiento continuo de la base de datos. Su complejidad ha ido en aumento, siendo su empleo rutinario menos amigable que otros modelos actuales (41).

En nuestro medio (España y Europa) el principal referente hasta la fecha ha sido EuroSCORE, en los últimos años en proceso de sustitución por EuroSCORE II, cuyas particularidades desarrollamos con más detalle a continuación.

1.3.1. EuroSCORE

EuroSCORE nació como proyecto en 1993 como respuesta europea a la escala de riesgo de Parsonnet, ya mencionada previamente (41). A diferencia del sistema de trabajo establecido por la base de datos de la STS, se construyó una base de datos *de novo* con las aportaciones de los centros que participaron en el proyecto. Las variables predictoras seleccionadas se basaron en los datos derivados de la literatura existente y en otros modelos predictivos previos. Del mismo modo se adoptó una definición de mortalidad postoperatoria como aquella

sucedida previamente al alta o en los 30 primeros días tras la intervención. Se elaboraron formularios en formato papel para la recogida de los datos, y la propuesta de participación se realizó a través de correo y mediante la invitación directa por parte de los miembros del comité de dirección del proyecto. La decisión final de participación de los centros fue voluntaria (9).

La recogida de datos se realizó entre los meses de septiembre y diciembre de 1995, y se introdujeron todos los pacientes intervenidos bajo circulación extracorpórea durante ese periodo en los centros participantes. Tras la depuración de datos se obtuvo un conjunto de 19.030 pacientes provenientes de 132 centros ubicados en ocho países europeos diferentes. Los datos fueron introducidos manualmente por dos personas distintas, y sometidos a revisión para la identificación de errores y registros incompletos. Para la elaboración del modelo se realizó inicialmente un análisis univariante, que consiste en calcular el grado de asociación estadística de cada variable con el evento de mortalidad por separado, utilizando en este caso un nivel de significación estadística de 0,2. Se decidió la inclusión de cada variable si se demostraba esta asociación estadística y además estaba presente en al menos el 2% de los pacientes. Posteriormente se realizó un análisis multivariante, que consiste en el cálculo del peso de cada variable dentro de la ecuación final mediante técnicas de regresión estadística, que conformó el modelo final de 17 variables.

Se incluyó un estudio de validación interna en el mismo conjunto de datos del estudio de desarrollo mediante sistema de división simple de la muestra. La calibración se evaluó calculando el área bajo la curva ROC (*Receiver Operating Characteristic*), con un valor de *c* de 0,76 (intervalo de confianza al 95% no comunicado) para el grupo de validación. La discriminación se evaluó mediante la prueba de bondad de ajuste de Hosmer-Lemeshow, con un resultado de 7,5 (p=0,68). Ambos resultados corroboraron una buena validez interna del modelo.

Este trabajo fue presentado y publicado en 1999. Se construyeron dos versiones diferentes; una logística, que calculaba la probabilidad de muerte de los pacientes

de acuerdo con los cálculos de una regresión logística binaria, y en la que se aplicó un coeficiente β a cada variable, y otra aditiva, en la que se otorgó un peso específico a cada una de las variables en relación con el coeficiente β de la regresión logística. La ventaja de esta última versión consistía en la extrema sencillez en el cálculo del riesgo quirúrgico a pie de cama del paciente sin necesidad de calculadora, y por esa causa fue la primera versión de EuroSCORE publicada (8). Los datos relativos a los coeficientes β derivados de los cálculos de la regresión logística no fueron publicados hasta pasados cuatro años de la primera publicación (42).

Los años que siguieron a la publicación de EuroSCORE se caracterizaron por la realización de múltiples estudios de validación externa y de análisis de subgrupos en la base de datos original. En este último aspecto destaca el estudio realizado por el mismo equipo de desarrollo de EuroSCORE, en el que analizaron los datos de los pacientes sometidos a cirugía coronaria en cada uno de los países mayoritarios participantes (aquellos que participaron con más de 500 pacientes). Este estudio es relevante porque es el único análisis de datos de la población española a nivel nacional que se había realizado hasta el momento, y arroja una serie de conclusiones interesantes, como que el perfil de riesgo, así como la mortalidad operatoria de los pacientes españoles (8,3%) es mayor cuando se comparan con otros países del norte de Europa (EuroSCORE medio 4,8%) (43). España participó en el desarrollo de EuroSCORE con los datos de 2.422 pacientes provenientes de 25 centros diferentes, de los cuales 1.119 habían sido sometidos a cirugía coronaria. Incluso la proporción de este tipo de cirugía fue sensiblemente menor (46,2%) que la comunicada por el resto de países (62,8% de media, con países como Alemania declarando hasta un 73,4% de cirugía coronaria) (44). A pesar de ello el modelo tuvo una adecuada validez interna dentro del subgrupo de pacientes españoles de la base de datos original (43).

Los estudios de validación externa fueron múltiples, y casi siempre con buenos resultados en la época inicial. Se obtuvieron buenos valores de calibración y discriminación en poblaciones geográficamente muy dispares, como en centros

de Finlandia, Brasil, Taiwán, Holanda o Cataluña, demostrándose además su eficacia en distintos grupos de patología (45-50). Un dato a reseñar es que, a pesar de que el diseño original de EuroSCORE no incluyó pacientes sometidos a cirugía coronaria sin circulación extracorpórea, varios estudios demostraron una adecuada capacidad de predicción para esta técnica quirúrgica (51-54).

Mención especial merece el gran estudio de validación externa de EuroSCORE en Norteamérica, con casi 600.000 pacientes extraídos de la base de datos de la STS. Esta validación se realizó mediante la división de la muestra en dos grupos diferentes, los intervenidos en 1995, que coincidían con el año de desarrollo de EuroSCORE (188.913 pacientes) y otro grupo de pacientes intervenidos en 1998 y 1999 (401.684 pacientes). En ambos grupos se obtuvieron adecuados parámetros de validación (55).

Este modelo puede ser catalogado como un éxito desde el mismo momento de su aparición debido a su enorme popularidad y a los excelentes resultados obtenidos inicialmente. A pesar de ello también existieron opiniones críticas y trabajos con resultados adversos en los primeros tiempos (56), de tal manera que en el año 2003 los creadores de EuroSCORE, realizaron una nueva validación del modelo en la base de datos original del desarrollo, obteniendo nuevamente buenos parámetros de calibración y discriminación tanto para la versión aditiva como para la logística (57).

Es a partir de la segunda mitad de la década de los 2000 cuando comienzan a aparecer estudios demostrando una pérdida de calibración de EuroSCORE. Una análisis de este proceso realizado con posterioridad mostró un progresivo aumento de la complejidad y perfil de riesgo de los pacientes sometidos a cirugía cardiaca, a la par que la mortalidad decreció de forma continua por la mejoría en las técnicas y habilidades quirúrgicas, así como por los avances en perfusión y cuidados postoperatorios (58).

Un importante estudio realizado en Australia con 8.331 pacientes mostró un fallo de calibración por sobreestimación del riesgo, tanto en el conjunto total de pacientes como en el subgrupo de cirugía coronaria. Aun así la discriminación resultó excelente, con áreas de la curva ROC por encima de 0,8 en ambos grupos (59). Similares resultados se obtuvieron en otro estudio realizado sobre una pequeña población de pacientes valvulares intervenidos en China (60).

Particularmente importantes son los malos resultados de calibración que mostró EuroSCORE en dos grupos de especial relevancia en la actualidad, como son los pacientes sometidos a cirugía valvular aórtica y los de alto riesgo quirúrgico, definiéndose estos últimos de forma variable según los estudios, aunque se acepta que son aquellos pacientes cuyo riesgo de muerte supera un tanto por ciento (7%-20%) (61, 62) o se sitúan en el tercil de mayor riesgo en las muestras analizadas (8). Algunos estudios ya habían comunicado previamente la existencia de sobreestimación del riesgo de mortalidad operatoria en ambos grupos (63, 64). La adecuada valoración del riesgo quirúrgico en estos pacientes resulta de especial importancia en la época actual, en la que las principales innovaciones tecnológicas están específicamente orientadas al tratamiento de estos subgrupos, como en el caso del implante transcatéter de prótesis aórticas o el implante de prótesis aórticas sin sutura. Son también de interés los estudios que han comparado el rendimiento de EuroSCORE con el de otros modelos de predicción del riesgo quirúrgico, principalmente el *STS Score*. Aunque algunos de estos estudios tienen un tamaño de muestra insuficiente (65), muestran una tendencia que sí es constatada en otros estudios posteriores más amplios (66, 67), como es la progresiva pérdida de calibración con sobreestimación del riesgo quirúrgico de EuroSCORE frente al resto. Enfatizando en este aspecto, durante el estudio PARTNER (*Placement of Aortic Transcatheter Valves*), se observó tal discrepancia entre los riesgos estimados por EuroSCORE y *STS Score*, que el umbral de mortalidad esperada para definir al paciente de alto riesgo quirúrgico fue diferente para cada uno de los modelos (61, 62).

1.3.2. EuroSCORE II

A finales de la década de los 2000 parecía clara y globalmente aceptado que la capacidad predictiva de EuroSCORE había derivado hacia una pérdida de calibración por sobreestimación del riesgo quirúrgico. Esto dio lugar a la aparición de diversos estudios más o menos individualizados de recalibración (68-70) y de análisis de los mecanismos de fallo del modelo (71). Aun así el hecho más importante que se derivó de esta situación fue la aceptación de la misma por los mismos autores de EuroSCORE, que diseñaron y pusieron en marcha el proyecto EuroSCORE II para la actualización del modelo de predicción (72).

Fruto de este esfuerzo, la versión actualizada de EuroSCORE fue presentada en el congreso de la EACTS *(European Association for Cardio-Thoracic Surgery)* celebrado en Lisboa en octubre de 2011. La publicación escrita del trabajo original tuvo lugar durante el primer trimestre de 2012.

El nuevo modelo fue desarrollado durante el año 2010. Se realizó una llamada global a la participación voluntaria para el desarrollo del mismo a través de anuncios en revistas especializadas, congresos, internet y correo electrónico. El interés generado fue máximo, con 214 centros registrados, de los cuales 154 completaron correctamente los datos. Como primera diferencia con el estudio que generó el primer modelo en el año 1995, los centros participantes no se redujeron solamente al ámbito europeo, sino que además se aportaron datos para análisis desde 43 hospitales de otros continentes, incluyendo centros americanos (Uruguay, Brasil, Canadá, Argentina, Estados Unidos…), africanos (Sudán, Sudáfrica…), de oriente medio (Siria), Asia (China, India, Japón, Taiwán…) e incluso Oceanía (Nueva Zelanda). Curiosamente, España fue el país del que más centros participaron en número (19), aunque probablemente no aportó el mayor volumen de pacientes (41).

Previamente a la recogida de datos se analizó la literatura en busca de campos de potencial mejora del antiguo modelo, siendo las variables más demandadas la

caracterización más precisa de la función renal preoperatoria por medio del aclaramiento de creatinina, la solicitud de la disfunción hepática como predictor y la redefinición de la angina inestable con criterios distintos a los de la versión previa.

La recogida de datos se realizó mayoritariamente de forma electrónica, en línea a través de un formulario alojado en una página de internet, aunque algunos grupos utilizaron como soporte el papel y otros remitieron sus propios datos en diferentes formatos, que fueron introducidos manualmente por los autores. Se recopilaron los datos de los pacientes intervenidos entre el 3 de mayo y el 25 de julio de 2010 (ambos inclusive, con un tiempo total de 12 semanas). Se registró el estatus postoperatorio (vivo o fallecido) en tres momentos diferentes: al alta hospitalaria, a los 30 días y a los 90 días de la intervención. Las variables solicitadas en el formulario están resumidas en la Tabla 1.

Tabla 1. Factores de riesgo recogidos para el desarrollo de EuroSCORE II.	
Factores relacionados con el paciente	
Edad	Arteriopatía extracardiaca
Sexo	Disfunción neurológica o musculoesquelética
Peso	Diálisis
Altura	Creatinina
Diabetes	BNP (péptido cerebral natriurético)
Albúmina sérica	
Factores dependientes del estado cardiológico	
Clase funcional NYHA	Infarto reciente y tamaño
Clase funcional CCS	Presión pulmonar sistólica
Disfunción sistólica	Endocarditis activa
Factores dependientes de la intervención	
Prioridad de la intervención	Tiempos de circulación extracorpórea
Tiempo de pinzamiento aórtico	Tiempo de parada circulatoria e hipotermia
Perfusión cerebral selectiva	Cirugía cardiaca previa
NYHA: *New York Heart Association*; CCS: *Canadian Cardiovascular Society*	

Se recogieron los datos de casi 25.000 pacientes, que tras análisis, depuración y eliminación de datos incompletos al azar, registros que no cumplían criterios de inclusión (válvulas transcatéter, trasplantes) y pacientes duplicados, se redujeron a 22.381 casos finales.

Para la elaboración del modelo se adoptó una aproximación basada en la parsimonia para lograr una interacción más simple y amigable con el cirujano. La muestra fue dividida en un grupo de desarrollo (75% de los pacientes) y un grupo de validación interna (25%) mediante muestreo aleatorio. Inicialmente se realizó un análisis univariante de los predictores para comprobar asociaciones, umbrales de riesgo y realizar una adecuada categorización de las variables continuas. El modelo inicial constaba de 14 variables, a las que se añadieron cuatro más mediante el método *forward*, comprobando una adecuada estabilidad y ajuste del conjunto en cada paso. Tras la realización de la regresión logística se obtuvieron los coeficientes β para cada variable y la constante de la ecuación, resultando en el modelo definitivo compuesto por 18 variables (una continua, seis discretas multicategóricas y 11 discretas dicotómicas).

La justificación final de la inclusión de ciertas variables, así como la exclusión de algunas otras, requirió un comentario adicional por parte de los autores. La disfunción hepática, que a pesar de algún resultado discordante (73) ha sido reconocida como factor predictivo independiente de mortalidad en múltiples trabajos (4-7), se intentó caracterizar por medio del valor de albúmina sérica. Desafortunadamente no se apreció relación entre este dato analítico y la mortalidad peroperatoria, probablemente debido a un problema en la recogida de datos originado por las distintas unidades de medida empleadas en los distintos laboratorios.

Otras variables, como la fracción de eyección del ventrículo izquierdo y la presión arterial pulmonar sistólica fueron divididas en más categorías que en el modelo previo. Finalmente no se incluyeron la rotura septal post-infarto ni el BNP *(Brain Natriuretic Peptide)* en el modelo por su escasa incidencia en el primer caso, y por falta de datos en el segundo (solo disponible en el 7,3% de los casos).

Se realizó una validación interna del modelo original mediante la evaluación de la calibración y la discriminación. Los resultados fueron buenos, especialmente la discriminación, con un área bajo la curva ROC de 0,81 (intervalo de confianza al

95% 0,78-0,84). La prueba de Hosmer-Lemeshow obtuvo un valor del estadístico de 15,9 y un valor de p de 0,0505, muy cercano al fallo de calibración. Este resultado fue descrito como una "ligera aunque aceptable tendencia a la infraestimación del riesgo quirúrgico".

Sin embargo, probablemente el punto más débil en el diseño de EuroSCORE II viene marcado por la definición de mortalidad postoperatoria empleada. Esta variable de resultado tuvo que ser redefinida por el escaso seguimiento obtenido en los datos aportados por los centros participantes, a pesar de que los autores se esforzaron en argumentar matemáticamente que la definición tradicional de muerte peroperatoria (intrahospitalaria o hasta 30 días tras la intervención) apenas aportaba un 0,6% extra de mortalidad observada a la obtenida considerando solamente el periodo de hospitalización. El seguimiento a 30 días fue tan solo del 56,6%, mientras que el seguimiento a 90 días cayó hasta el 44,4%. Este hecho obligó a redefinir la variable resultado como mortalidad estrictamente intrahospitalaria, y es probablemente el aspecto más duramente criticado de este modelo.

Por último es conveniente resaltar que una de las hipótesis de este trabajo de validación del modelo fue mencionada como una de las debilidades del diseño de EuroSCORE II. El sesgo inducido por la participación voluntaria pudo ocasionar que el aporte de datos se produjera mayoritariamente desde centros con buenos resultados (habitualmente grandes centros de alto volumen), a pesar de la explícita llamada a la participación de centros de cualquier tipo, volumen y resultados, por parte de los autores (10).

Las críticas al modelo no se hicieron esperar, y ya la misma publicación vino acompañada por un duro editorial que criticaba varios aspectos de EuroSCORE II (11). En primer lugar se criticó la elección de variables y se demandó el desarrollo de modelos por patología, arguyendo que existen variables de especial importancia en algunos grupos de pacientes que pueden no serlo para otros casos. También fue duramente criticado en este editorial la escasa renovación de

variables, todas muy similares a las empleadas en EuroSCORE, excluyendo nuevos conceptos que actualmente se valoran como importantes, especialmente en el paciente anciano (fragilidad, estatus mental, etc.), así como el tratamiento matemático de las variables continuas. Asimismo se criticó, como se mencionó previamente, que se hubiese considerado adecuado el resultado arrojado por la prueba de Hosmer-Lemeshow realizada en el grupo de validación interna del original. Sin embargo, y como era previsible, el aspecto más duramente criticado fue la definición de mortalidad como exclusivamente intrahospitalaria. El autor de esta crítica definió la base de datos como escasa por el limitado seguimiento del estatus de los pacientes, reprochando a las unidades que participaron en el estudio aportando los datos su falta de implicación, e incluso de ética profesional. El modelo fue desaconsejado para su uso como monitorización de la calidad asistencial o información al paciente.

Hubo otros que remarcaron posibles problemas en el desarrollo del modelo, criticándolo fundamentalmente por su tendencia a la infraestimación del riesgo quirúrgico. Resulta curiosa incluso la existencia de un artículo que describió la existencia un ritmo circaanual en la mortalidad peroperatoria en Cirugía Cardiovascular (74), y cómo la recogida de datos para el desarrollo de EuroSCORE II se realizó precisamente en la época del año en que se observa una menor mortalidad quirúrgica (75). Siguiendo el hilo del editorial mencionado, se reclamaron explicaciones del resultado de la prueba de bondad de ajuste, dentro de la validación interna en el estudio original de EuroSCORE II. Especialmente importante fue la publicación de una carta al editor en la que se solicitaba una explicación de este hecho al equipo de desarrollo, así como una aclaración de la metodología empleada para realizar la validación interna, ya que la redacción en el artículo era bastante confusa (76). El grupo creador del modelo respondió con datos no publicados en el estudio original, en el que aportaban otro valor diferente para la prueba de Hosmer-Lemeshow (14,9, con un valor de p de 0,09), concluyendo que los parámetros de calibración fueron realmente satisfactorios (77). Este mismo problema metodológico fue también comentado en otro artículo posterior (78), que recibió similar respuesta (79).

A pesar de las dudas generadas, este rediseño de EuroSCORE había sido largamente esperado, y no tardaron en aparecer múltiples estudios de rendimiento y validación externa. Una de las primeras consideraciones que se realizaron en el original de EuroSCORE II fue que la probabilidad media de muerte peroperatoria era inferior a la calculada por EuroSCORE para el mismo grupo de pacientes. Este punto fue desarrollado en un estudio que demostró que, como media, el valor de EuroSCORE II era inferior al de EuroSCORE, por lo que se consideraba cumplido el objetivo de corrección de la sobreestimación de la mortalidad (80).

Desde esa fecha hasta el momento actual se han publicado diversos estudios de validación con resultados contradictorios sobre el rendimiento del nuevo modelo. El estudio realizado por Biancari et al. en pacientes coronarios en un centro finlandés mostró una adecuada calibración y discriminación de EuroSCORE II (81), aunque un estudio restrospectivo realizado por Di Dedda et al. en un único centro de Italia mostró una tendencia a la infraestimación del riesgo de mortalidad, especialmente en el subgrupo de pacientes de alto riesgo quirúrgico (82).

En un estudio más amplio realizado por Chalmers et al. en una base de datos de pacientes británicos los resultados fueron menos optimistas, con datos inconsistentes, que reflejaban buena calibración en algunos subgrupos, a la vez que un muy pobre rendimiento en otros, especialmente en el grupo de pacientes sometidos a cirugía valvular aórtica (83).

Otro estudio de validación realizado por Zhang et al. en población china mostró un rendimiento extraordinariamente bajo de EuroSCORE II, mostrando la ya mencionada tendencia a la infraestimación del riesgo quirúrgico. Curiosamente este estudio mostró bajo rendimiento incluso en términos de discriminación, que en trabajos previos había sido consistentemente excelente (84). Otro estudio más amplio de los mismo autores volvió a mostrar valores de calibración y capacidad de discriminación muy pobres (85). Algunos trabajos más modestos en cuanto a

tamaño de la muestra analizada mostraron fallos de calibración en la población turca, argentina y pakistaní (30, 86, 87).

Por el contrario, recientemente han surgido otros trabajos de validación externa que sí obtiene adecuados parámetros de rendimiento. Un artículo de un grupo italiano firmado por Paparella et al. obtuvo muy buena calibración y discriminación en pacientes sometidos a cirugía coronaria con y sin circulación extracorpórea, aunque con la ya conocida tendencia a la infraestimación en el grupo de mayor riesgo quirúrgico (88). Uno de los trabajos más importantes publicados hasta la fecha fue el realizado con la base de datos de la STS con más de 50.000 pacientes. Sus conclusiones mostraron que *STS Score* muestra mejores resultados en esa población, lo cual es lógico, ya que se trataría de una validación interna, donde siempre existe cierto grado de optimismo, aunque EuroSCORE II fue considerado una buena alternativa, especialmente en cirugía coronaria y de bajo riesgo, en este trabajo definido como mortalidad estimada de hasta el 5% (89).

Para aportar algo más de caos a la situación actual en cuanto a la fiabilidad de EuroSCORE II, varios estudios multicéntricos realizados en Italia mostraron varios aspectos interesantes en cuanto al rendimiento de la nueva escala; en primer lugar, un primer estudio publicado por Barili et al. mostró muy buen rendimiento en términos de calibración en pacientes de bajo riesgo sometidos a cirugía valvular aórtica, aunque se apreció fallo por infraestimación del riesgo en los terciles de medio y alto riesgo (90). Este trabajo presenta el defecto de contar con una muestra reducida, aunque los mismos resultados fueron reproducidos en otro estudio de los mismos autores, esta vez realizado con un tamaño de muestra 10 veces mayor (91). De nuevo el mismo equipo realizó un nuevo análisis en otro subgrupo de pacientes sometidos a cirugía valvular mitral, que concluyó de la misma forma: buena capacidad de discriminación y adecuada calibración en pacientes de bajo riesgo, siendo inadecuada por infraestimación en pacientes de medio y alto riesgo (92).

Anecdóticamente, podemos llegar a encontrar trabajos (de escaso tamaño y relevancia) que defienden que la simple evaluación clínica y el buen criterio del cirujano conforman un modelo predictivo tan consistente y con tan buenos parámetros de rendimiento como EuroSCORE, EuroSCORE II o *STS Score* (93).

Dentro de las distintas corrientes de opinión existentes entre los autores de referencia en estudios de Calidad acerca del buen rendimiento de EuroSCORE, quizá una de las principales sugerencias es la de adoptar un método de trabajo de inclusión constante e ininterrumpida de datos, junto con la de recalibrar el modelo de forma periódica, utilizando la metodología que emplea *STS Score* (94). La respuesta del grupo de trabajo que desarrolló el modelo sigue siendo de defensa del mismo, aun a pesar de las deficiencias en el seguimiento de pacientes y de los contradictorios resultados de los estudios de validación realizados hasta el momento (95, 96).

Dentro de nuestro país, existen dos estudios con resultados muy consistentes en cuanto a rendimiento externo de EuroSCORE II. El primero por orden cronológico fue un estudio retrospectivo realizado en un solo centro publicado por Silva et al. (97, 98), que mostró una mortalidad observada situada entre los valores pronosticados por EuroSCORE y EuroSCORE II, con fallo de calibración y buena discriminación por parte de ambos modelos. Este estudio fue criticado, resaltándose particularmente el hecho de que los valores medios de mortalidad esperada eran muy altos si se comparaban con otras publicaciones existentes en la literatura (41). El segundo estudio de relevancia realizado hasta el momento en nuestro país ha sido publicado por García-Valentín el al. (99, 100), y se trata de un estudio prospectivo y multicéntrico que obtuvo datos muy similares al anterior, con mortalidades esperadas muy altas, probablemente reflejo de la gran complejidad y prevalencia de comorbilidades que presentan los pacientes intervenidos en nuestro país, y con un resultado de mortalidad observada situado entre los pronosticados por los dos modelos, quizá algo más cercano a EuroSCORE II, pero con fallo de calibración en ambos y conservando una buena capacidad de discriminación.

1.4. Relación volumen-resultado

Podemos definir como volumen de una actividad el número de veces que se repite esa intervención o procedimiento en un hospital, servicio o por un cirujano individual. Este concepto ha sido utilizado como un marcador de calidad de forma usual, a pesar de que la relación entre volumen y resultado es compleja de evaluar y definir.

Existen diversos estudios que analizan esta relación. En el estudio de Clark sobre la base de datos de la STS de 1996 (101) se demostró que existía una relación entre volumen y resultados, aunque difusa y que no obedecía a una progresión lineal o geométrica, ni lograba esclarecer un punto de corte (en términos de volumen) a partir del cual se determinaba un nivel de seguridad estable. Las causas descritas para este fenómeno son muchas, comenzando por el manejo de datos de mortalidad cruda, y no ajustada al riesgo (pueden existir centros de bajo volumen y mortalidad ajustada alta, pero que en términos absolutos sea menor que la de otros centros que manejen pacientes extremadamente complejos y tengan mortalidad ajustada baja). Asimismo se menciona la importancia del cirujano individual y su entorno hospitalario, que plantea que los resultados individuales de cada cirujano se ven ampliamente afectados por la calidad asistencial del hospital en el que desempeña su tarea. La conclusión más importante de este estudio fue, sin duda, la de la necesidad de estratificar de forma adecuada el riesgo quirúrgico.

El estudio en el que se halló una mayor evidencia de la asociación inversa entre volumen y mortalidad quirúrgica fue el publicado por Birkmeyer et al. en 2002. Este trabajo, que se realizó con datos extraídos de la base de datos de *Medicare* (el sistema de cobertura de seguridad social financiado por el gobierno estadounidense), tuvo la particularidad de comparar las tasas de mortalidad de acuerdo a una estratificación de los pacientes en base a edad, sexo, etnia, comorbilidades y prioridad de la intervención, realizando un ajuste al riesgo por grupos. Se analizó la relación volumen-mortalidad en 14 intervenciones

cardiovasculares u oncológicas, entre las que se incluyeron la cirugía coronaria y la valvular aislada. En ambos casos hubo una disminución constante de la mortalidad peroperatoria observada con cada estrato de volumen de actividad (muy bajo, bajo, medio, alto, muy alto, definidos por división en quintiles de la muestra). La *odds ratio* para la mortalidad entre hospitales de muy alto y muy bajo volumen, varió entre 0,75 y 0,8 para cada tipo de intervención (reducción de la tasa relativa de mortalidad del 20 al 25%) (102). Estos datos fueron confirmados en otro estudio realizado por Dudley et al. en California, en el que además se defendía la política de referencia a centros de alto volumen con el objetivo de ahorrar muertes evitables y atribuibles al bajo volumen de un centro (103). La relación entre nivel de actividad y resultados quirúrgicos también ha sido demostrada recientemente en el ámbito de la cirugía cardiaca pediátrica (104).

Otros estudios publicados realizaron interesantes matizaciones acerca de esta relación, como el publicado por Papadimos et al. en 2005, que mostró similares cifras de mortalidad y complicaciones en un centro de bajo volumen, comparado con la predicción del *STS Score*. La particularidad de este estudio reside en que la intervención y la protocolización de los cuidados intensivos postoperatorios eran realizados por cirujanos experimentados provenientes de hospitales de alto volumen, importando su actividad y protocolos de actuación (105). Otro estudio realizado por Marcin et al. en California y publicado en 2008, apreció la mencionada relación entre una mayor actividad quirúrgica y mejores resultados. Sin embargo ésta desapareció posteriormente, cuando se hizo obligatoria la declaración de resultados en dicho estado (106). Dicho fenómeno se apreció igualmente en el estado de Nueva York, cuando en los años 90 se hizo obligatoria la declaración de mortalidad por hospital y cirujano en dicho estado. Este hecho consiguió una reducción relativa de la mortalidad del 41% en cuatro años (107). En ambos casos la reducción de las cifras de mortalidad operatoria se atribuyeron a la mejoría de los resultados en los centros y cirujanos de bajo volumen, ante la obligatoriedad de publicación de su mortalidad de forma individual. A pesar del evidente efecto beneficioso de la publicación de resultados en términos de disminución de la mortalidad existen potenciales consecuencias negativas de

esta práctica, como los cambios en los volúmenes de casos complejos de unos centros a los otros o los cambios en las prácticas de los cirujanos (selección de casos de bajo riesgo, altas rutinarias de pacientes no candidatos al alta hospitalaria, sobrecodificación de factores de riesgo, etc.) (37).

Además de evaluar la relación entre práctica individual y resultados, existen estudios que aprecian un aumento de la mortalidad individual de un cirujano cuando realiza intervenciones de similar riesgo en centros de menor volumen, concluyendo en que también existe una relación volumen-mortalidad dependiente del centro, enfocada hacia la calidad del proceso asistencial postoperatorio (108).

Aun así los resultados de los diferentes estudios muestran conclusiones complejas y en ocasiones contradictorias. Podemos encontrar otros trabajos en que no se apreció ninguna diferencia entre la mortalidad peroperatoria de los centros de menor y mayor volumen, especialmente en el campo de la cirugía coronaria (109), así como estudios en los que existe una evidente relación entre actividad y mortalidad, aunque compleja y sin poderse excluir los buenos resultados en centros de pequeño volumen (110).

Como resumen, parece existir una relación entre un alto volumen de actividad de un centro y una baja mortalidad quirúrgica, aunque compleja en cuanto a su definición y cuantificación. Ésta depende además de múltiples factores, como la actividad individual del cirujano, la calidad de los protocolos de cuidado postoperatorio y los recursos existentes (111, 112). Por otro lado, resulta imprescindible realizar una adecuada estratificación del riesgo quirúrgico con modelos predictivos adecuadamente validados en el medio en que se realiza la actividad para la evaluación de los centros.

1.5. Justificación del estudio

La estratificación del riesgo quirúrgico obedece a una necesidad inherente a la práctica médica, que es la de conocer qué pacientes se beneficiarán de nuestra actuación y cuáles no lo harán o incluso serán perjudicados por ella,

contraviniendo así el principio de no maleficencia. La creación de herramientas que permiten estimar la probabilidad de mortalidad o de complicaciones en los pacientes que se someten a Cirugía Cardiovascular son útiles para realizar con máxima precisión las indicaciones quirúrgicas, dar una correcta información a los pacientes para que sean capaces de decidir razonadamente si someterse o no a la intervención quirúrgica propuesta y dotar de datos comparativos para la evaluación de la calidad del la actividad quirúrgica desarrollada (113).

La elaboración de estas herramientas se basa en la construcción de modelos predictivos mediante métodos que utilizan el grado de asociación estadística entre algunas variables predictoras y el resultado final cuya probabilidad de aparición quiere ser estimada, otorgando a cada variable un valor o peso específico para su cálculo. Las escalas de estimación de riesgo quirúrgico empleadas en Cirugía Cardiovascular son modelos predictivos clínicos aplicados específicamente a nuestro campo de actuación. Este campo ha presentado un importante desarrollo, y una simple búsqueda bibliográfica arroja miles de referencias de trabajos publicados en los últimos años sobre este tema (2).

La Cirugía Cardiovascular ha sido una de las especialidades médicas más estudiadas a lo largo de su historia. Como ejemplo, se ha constatado que la cirugía coronaria o de revascularización miocárdica ha sido el tratamiento más estudiado, auditado y contrastado de la historia de la Medicina, junto a la herniorrafia inguinal y la colecistectomía (113-115). El empleo de modelos de predicción del riesgo quirúrgico entra dentro de la práctica habitual del cirujano cardiovascular desde hace años. Este hecho, junto a la existencia de una inquietud inherente por la recogida de datos y el análisis de resultados, la convierten probablemente en una de las especialidades con más capacidad de generar registros de actividad y calidad en el mundo médico.

Existen multitud de modelos de estimación del riesgo quirúrgico elaborados a lo largo de la historia de la Cirugía Cardiovascular (39, 116-118), algunos incluso desarrollados en España (119), que a día de hoy se encuentran en su mayoría en

desuso. Los modelos más empleados en la actualidad son el publicado por la *Society of Thoracic Surgeons* norteamericana (*STS Score*) y el EuroSCORE (*European System for Cardiac Operative Risk Evaluation*), de origen europeo. El *STS Score* es revisado y ajustado periódicamente, habiéndose efectuado su última actualización en 2009. Predice el riesgo de mortalidad y complicaciones mayores peroperatorias intrahospitalarias (insuficiencia renal aguda, accidente cerebrovascular, ventilación mecánica prolongada, infección profunda de la herida esternal y reoperación) y a 30 días, así como un evento combinado de morbimortalidad (120). EuroSCORE es un modelo creado a partir de datos de pacientes europeos que fueron recogidos en 1995. Este modelo predice mortalidad intraoperatoria, postoperatoria intrahospitalaria y a 30 días, y ha sido el más empleado en nuestro país desde su aparición en 1999 (8, 9).

Sin embargo en los últimos años de la década pasada, EuroSCORE ha mostrado una serie de fallos de calibración en varios estudios, apareciendo una progresiva tendencia a la sobreestimación del riesgo quirúrgico, especialmente en cirugía valvular (60, 63, 71). Esta tendencia encaja con la discrepancia observada entre la mortalidad estimada por el *STS Score* y EuroSCORE para la cirugía valvular aórtica en el contexto de estudios de crucial importancia, como en los PARTNER (61, 62), trabajos de gran impacto en la práctica actual a pesar de la limitación metodológica que presentan al tratarse de estudios de no inferioridad. Es por todo esto que en 2010 se inició un nuevo proyecto de recogida de datos que concluyó con la publicación en 2012 de EuroSCORE II, una versión actualizada del antiguo modelo (10). En ella se han revisado y redefinido alguna de las antiguas variables, se ha eliminado la rotura septal post-infarto como predictor independiente por su escasa prevalencia y se han añadido algunas variables nuevas. La mortalidad peroperatoria pronosticada por este modelo es globalmente más baja, lo cual corrige los fallos de sobreestimación observados previamente.

El diseño del proyecto de desarrollo de EuroSCORE II se basó en la participación voluntaria de los centros interesados (europeos o no), que debían declarar los

datos preoperatorios y sobre el procedimiento realizado de todos los pacientes intervenidos durante los meses de mayo, junio y julio de 2010, así como el estado postoperatorio (vivo o fallecido) al alta hospitalaria, 30 y 90 días. Este diseño otorgó mayor relevancia, por número de pacientes incluidos y el diseño de la recogida de datos, a los datos de aquellos hospitales de mayor volumen de actividad. Los datos disponibles de los estudios de validación externa realizados hasta la fecha son dispares, aunque aparecen con frecuencia comunicaciones de fallos de calibración por infraestimación el riesgo quirúrgico del modelo.

La relación entre el volumen hospitalario o individual y la calidad de los resultados ha sido demostrada en numerosos estudios (101, 105, 109). Esta relación es sin embargo compleja, ya que los resultados de algunos procedimientos se han relacionado con el volumen institucional, y los de otros con los volúmenes individuales de los médicos responsables.

Dado que los resultados de mortalidad estimada según las predicciones de EuroSCORE II suponen un aumento en cuanto a las exigencias del estándar de calidad al ser utilizado como referencia para el cálculo de la misma, y la presumible mayor presencia de pacientes procedentes de centros de alto volumen, cabe la posibilidad de que el nuevo modelo no tenga un buen rendimiento en centros con menor volumen.

2. HIPÓTESIS DE TRABAJO

Existen numerosos modelos de predicción del riesgo quirúrgico en Cirugía Cardiovascular, siendo EuroSCORE II el más novedoso y actualizado en nuestro medio. Este modelo ha comenzado a generar estudios de validación externa con resultados contradictorios.

La relación volumen-resultado en Cirugía Cardiovascular es compleja y difícil de definir. Existen diversos trabajos que corroboran la existencia de esta relación, concluyendo que a mayor número de procedimientos realizados por un cirujano o centro (volumen) mejores son los resultados obtenidos (expresados en forma de menor mortalidad).

La primera versión de EuroSCORE es actualmente considerada un modelo obsoleto por múltiples estudios, en su mayoría considerando grandes series que incluyen centros de alto volumen. El diseño del proyecto de desarrollo de EuroSCORE II, en el que se declararon los resultados de los pacientes intervenidos en un intervalo cerrado de tiempo, tiene el problema metodológico de incluir un mayor número de pacientes aportados por centros de alto volumen y presumiblemente menor mortalidad, estando menos representados los centros de menor volumen. Esto llevaría, por tanto, a un escaso rendimiento de EuroSCORE II como modelo predictivo del riesgo quirúrgico cuando se aplica a los resultados de un centro de bajo o medio volumen, en el que infraestimaría la mortalidad.

La hipótesis nula (H_0) de este estudio supone que no existen diferencias estadísticamente significativas entre los resultados esperados por el modelo de predicción de riesgo quirúrgico EuroSCORE II y los observados para la misma población en un centro de medio volumen, validando el modelo para este tipo de centros.

La hipótesis alternativa (H_1) de este estudio supone que existen diferencias estadísticamente significativas entre los resultados esperados en base al modelo

de predicción de riesgo quirúrgico EuroSCORE II y los observados para la misma población en un centro de medio volumen, demostrando bajo rendimiento del modelo para este tipo de centros.

Se propone un estudio de validación externa que evaluará la calibración y discriminación de EuroSCORE II en un centro de medio volumen.

3. OBJETIVOS DEL TRABAJO

El objetivo primario es realizar una validación externa del modelo de predicción de riesgo quirúrgico en Cirugía Cardiovascular EuroSCORE II en un centro de medio volumen, el Hospital General Universitario de Alicante.

Los objetivos secundarios son:

1. Comprobar la adecuación de los resultados del Servicio de Cirugía Cardiovascular del Hospital General Universitario de Alicante a los estándares de calidad actuales, según la metodología aplicada en el Primer Proyecto de Calidad en Cirugía Cardiovascular de la Sociedad Española de Cirugía Torácica-Cardiovascular (121).
2. Realizar un análisis descriptivo de los factores de riesgo de la población del estudio.

3.1. Variables principales del estudio

- Valor medio del resultado de EuroSCORE II en la muestra de pacientes analizada.
- Mortalidad peroperatoria intrahospitalaria observada tras una intervención de cirugía cardiaca mayor (criterio de EuroSCORE II).

3.2. Variables secundarias del estudio

Están resumidas en las Tablas 2 y 3.

Tabla 2. Variables contenidas en EuroSCORE (tipo de variable, medida).	
Edad (continua, media +/- desviación estándar)	
Sexo (discreta dicotómica, %)	
Enfermedad pulmonar crónica (discreta dicotómica, %)	
Arteriopatía extracardiaca (discreta dicotómica, %)	
Disfunción renal (discreta dicotómica, %)	
Disfunción neurológica (discreta dicotómica, %)	
Endocarditis activa (discreta dicotómica, %)	
Cirugía cardiaca previa (discreta dicotómica, %)	
Situación preoperatoria crítica (discreta dicotómica, %)	
Disfunción sistólica (discreta multicategórica transformada en 3 variables dicotómicas excluyentes, %)	Fracción de eyección del ventrículo izquierdo > 50%
	Fracción de eyección del ventrículo izquierdo 30-50%
	Fracción de eyección del ventrículo izquierdo < 30%
Hipertensión pulmonar grave (discreta dicotómica, %)	
Angina inestable (discreta dicotómica, %)	
Infarto de miocardio reciente (discreta dicotómica, %)	
Cirugía urgente (discreta dicotómica, %)	
Cirugía distinta de coronaria aislada (discreta dicotómica, %)	
Cirugía sobre la aorta torácica (discreta dicotómica, %)	
Rotura septal post-infarto (discreta dicotómica, %)	

Tabla 3. Variables contenidas en EuroSCORE II (tipo de variable y medida).	
Edad (continua, media +/- desviación estándar)	
Sexo (discreta dicotómica, %)	
Enfermedad pulmonar crónica (discreta dicotómica, %)	
Arteriopatía extracardiaca (discreta dicotómica, %)	
Disfunción renal (discreta multicategórica, transformada a 4 variables dicotómicas excluyentes, %)	Aclaramiento de creatinina > 85 mL/min
	Aclaramiento de creatinina 51-85 mL/min
	Aclaramiento de creatinina < 51 mL/min
	Hemodiálisis crónica
Movilidad limitada (discreta dicotómica, %)	
Endocarditis activa (discreta dicotómica, %)	
Cirugía cardiaca previa (discreta dicotómica, %)	
Diabetes mellitus insulín-dependiente (discreta dicotómica, %)	
Situación preoperatoria crítica (discreta dicotómica, %)	
Disfunción sistólica (discreta multicategórica, transformada a 4 variables dicotómicas excluyentes, %)	Fracción de eyección del ventrículo izquierdo > 50%
	Fracción de eyección del ventrículo izquierdo 31-50%
	Fracción de eyección del ventrículo izquierdo 21-30%
	Fracción de eyección del ventrículo izquierdo < 21%
Hipertensión pulmonar (discreta multicategórica, transformada a 3 variables dicotómicas excluyentes, %)	Presión arterial pulmonar < 31 mmHg
	Presión arterial pulmonar 31-55 mmHg
	Presión arterial pulmonar > 55 mmHg
Angina de reposo (discreta dicotómica, %)	
Infarto de miocardio reciente (discreta dicotómica, %)	
Prioridad de la intervención (discreta multicategórica, transformada a 4 variables dicotómicas excluyentes, %)	Intervención electiva
	Intervención preferente
	Intervención urgente
	Intervención de salvamento
Complejidad de la intervención (discreta multicategórica, transformada a 4 variables dicotómicas excluyentes, %)	Cirugía coronaria
	1 procedimiento no coronario
	2 procedimientos
	3 o más procedimientos
Cirugía sobre la aorta torácica (discreta dicotómica, %)	
Clase funcional preoperatoria según la *New York Heart Association* (discreta multicategórica, transformada a 4 variables dicotómicas excluyentes, %)	NYHA I
	NYHA II
	NYHA III
	NYHA IV

4. MÉTODOS

4.1. Diseño del estudio

El estudio tiene un diseño longitudinal y ambispectivo. El periodo de inclusión de paciente se dividió en dos partes; una primera, previa al inicio del estudio, cuya recolección de datos se llevó de forma retrospectiva, y una segunda, que se inicia a partir del comienzo del estudio, en que ésta se realizó de forma prospectiva.

Se recopilaron los datos preoperatorios necesarios para el cálculo de EuroSCORE y EuroSCORE II de todos los pacientes que cumplían criterios de inclusión y fueron intervenidos en el centro del estudio durante el periodo calculado. Asimismo se registró la mortalidad peroperatoria intrahospitalaria de los pacientes incluidos en el estudio.

La unidad asistencial en que se realizó este estudio fue el Servicio de Cirugía Cardiaca del Hospital General Universitario de Alicante. Este centro pertenece a la red de hospitales públicos de la Agencia Valenciana de Salud y depende de la *Conselleria de Sanitat* de la *Generalitat Valenciana*. Durante el periodo de estudio el Servicio estaba compuesto por un Jefe de Servicio, cinco cirujanos de plantilla y dos médicos internos residentes, además de 3 perfusionistas y enfermería especializada tanto en quirófano (3 personas en cada intervención) como en las distintas áreas de cuidados postoperatorios (con una razón de un enfermero por cada dos pacientes en la unidad de críticos y por cada 15 pacientes en la sala de hospitalización). El control postoperatorio inmediato se realizaba principalmente en la Unidad de Reanimación, dependiente del Servicio de Anestesiología, Reanimación y Terapéutica del Dolor, siendo una minoría de pacientes manejados por el Servicio de Medicina Intensiva en su unidad. Los protocolos quirúrgicos y anestésicos estaban adecuadamente definidos y estandarizados, y los pacientes eran evaluados en sesiones médico-quirúrgicas, tomándose las decisiones de actuación en cada caso de forma consensuada entre los componentes del

Servicio. Todos los pacientes incluidos firmaron un consentimiento informado para la realización de la intervención propuesta.

La actividad promedio de esta unidad había sido de entre 300 y 350 intervenciones anuales de Cirugía Cardiaca mayor, tal y como se define posteriormente en los criterios de inclusión de pacientes, hasta la fecha del estudio. La calificación de centro de bajo, medio o alto volumen no está estandarizada, y es habitual que varíe en la literatura dependiendo de la distribución del número de intervenciones por unidad quirúrgica (101, 102). Habitualmente un centro con un número de intervenciones anuales similar al del Hospital General Universitario de Alicante suele ser catalogado de medio o bajo volumen en las series norteamericanas (102). Según datos europeos de la década de los 90, el volumen medio por centro en España era entonces de 266 pacientes por año (122). El último registro nacional publicado hasta el momento por la SECTCV (Sociedad Española de Cirugía Torácica-Cardiovascular) contiene los datos de actividad del año 2013, de los que es posible calcular una media de 366 intervenciones anuales por centro (123). Considerando el conjunto de estos datos de actividad, tanto nacionales como internacionales, hemos definido el Servicio en el que se realizó el estudio como un centro de medio volumen en nuestro entorno. También es importante destacar que el número de intervenciones por cirujano no es elevado (entre 50-60 por año), por lo que se cumple la premisa de que tanto el Servicio como los cirujanos realizan una actividad media.

4.2. Cálculo del tamaño de la muestra

El tamaño de la muestra necesario para el diseño del estudio de validación de un modelo predictivo depende fundamentalmente del número de eventos (en este caso, de muertes peroperatorias) que se estima que ocurrirán durante el periodo del estudio. El intervalo de confianza de la curva ROC es tanto más estrecho y preciso cuanto mayor es el tamaño de muestra. La prueba de bondad de ajuste necesita también tamaños de muestra grandes, aunque se discute cuál es el mejor número, aconsejándose que sea superior a 2.000 pacientes, aunque no

mayor de 3.000-4.000 (3, 25). Se ha estimado que un número de 50 eventos es suficiente para el cálculo de la curva ROC, aunque el tamaño de la muestra se ha ampliado ligeramente para obtener una prueba de bondad de ajuste más preciso, acorde a las recomendaciones de trabajos previos (3).

Se calculó un periodo de inclusión de pacientes de cuatro años, con un promedio de entre 300 y 325 pacientes por año, excluyendo los incluidos en el desarrollo de EuroSCORE II y esperando una mortalidad similar a la de otros estudios españoles (99, 100, 121, 124). La estimación fue que se podrían registrar entre 90 y 100 fallecimientos postoperatorios, que proporcionarían un tamaño suficiente para obtener una potencia estadística adecuada para la prueba de calibración.

4.3. Criterios de inclusión de pacientes

Los pacientes incluidos fueron todos aquellos pacientes adultos intervenidos en el periodo de estudio en el Servicio de Cirugía Cardiovascular del Hospital General Universitario de Alicante, a los que se les realizó una intervención de cirugía cardiaca mayor, con o sin circulación extracorpórea, definida ésta como aquella intervención quirúrgica realizada para el tratamiento de patología estructural del corazón y/o grandes vasos torácicos que precisa apertura de la caja torácica.

4.4. Criterios de exclusión de pacientes

Pacientes menores de 17 años, pacientes sometidos a cirugía cardiaca menor (refiriéndose en este centro principalmente a la cirugía de extracción de marcapasos sin circulación extracorpórea), cirugía de la pared torácica, implante de asistencias circulatorias, trasplante cardiaco o implante de válvulas transcatéter, al no estar incluidos estos procedimientos en el conjunto de pacientes de desarrollo de EuroSCORE II. También fueron excluidas las reintervenciones realizadas sobre un mismo paciente antes del alta, calculándose el riesgo de mortalidad del global del ingreso como el de la primera intervención.

Por otro lado fueron excluidos todos los pacientes que formaron parte de la aportación de datos que el Servicio realizó entre mayo y julio de 2010 para el proyecto de desarrollo de EuroSCORE II, ya que su presencia supondría que la validación no sería completamente externa.

4.5. Recogida de datos

4.5.1. Herramientas informáticas

La recogida de datos se realizó en hojas de cálculo realizadas mediante el programa Microsoft Excel 2011 ® (Microsoft Corporation, Redmond, WA, EEUU). El diseño de estas hojas de cálculo fue realizado por el autor del estudio, basándose en la calculadora fuera de línea de EuroSCORE, disponible en la página de internet www.euroscore.org (125). Esta hoja se adecuó para el cálculo de EuroSCORE II mediante la modificación de las variables (adición, eliminación o conversión a multicategórica) de acuerdo con las modificaciones descritas en el manuscrito original de EuroSCORE II (10). Asimismo se aplicó a cada variable el nuevo valor de coeficiente β e intercepto que se describían en el mismo. Esta hoja de cálculo fue comprobada, corroborando el adecuado funcionamiento de cada variable por separado y en conjunto, comparando los valores de la nueva herramienta con los de la calculadora en línea de EuroSCORE II, disponible también en la página de internet antes mencionada (125).

Se diseñaron dos hojas de cálculo adicionales, una para la detección de errores de coherencia basándose en variables lógicas y otra para volver a calcular los valores de EuroSCORE II tras la realización de modificaciones posteriores. Se realizó una búsqueda de errores en las variables multicategóricas (detectando si se habían marcado más de una categoría por variable), errores de coherencia entre variables de EuroSCORE y EuroSCORE II con idéntica definición (edad, sexo, arteriopatía, etc.) y de concordancia entre variables similares redefinidas en EuroSCORE II (angina de reposo y movilidad limitada). Se comprobaron los datos de todos los pacientes del estudio, y si se detectó algún fallo de coherencia se

recurrió de nuevo a la historia clínica para averiguar el valor correcto de la variable. En todos los pacientes que necesitaron revisión se utilizó la herramienta para el nuevo cálculo de EuroSCORE y EuroSCORE II.

Una vez los datos fueron completados, comprobados y depurados se volcaron a un archivo de IBM SPSS 20.0 ® (International Business Machines Corporation, Armonk, NY, EEUU) para su análisis estadístico.

4.5.2. Fase retrospectiva

Los datos de los pacientes intervenidos en 2010 y 2011, a excepción del peso y valor de creatinina preoperatorios, fueron recogidos de las historias clínicas hospitalarias, del programa hospitalario para la edición de informes de alta de la *Conselleria de Sanitat* de la *Generalitat Valenciana* (Alta Hospitalaria), y del programa de información ambulatoria de la misma entidad, Sistema de Información Ambulatoria. Los datos restantes (peso y creatinina basal) se obtuvieron de las hojas de perfusión de cada paciente. El estado de fallecimiento postoperatorio se extrajo de los mismos registros indicados además de la base de datos de actividad propia del Servicio de Cirugía Cardiovascular del Hospital General Universitario de Alicante. La recopilación de los datos de todas las variables de cada paciente de este periodo, así como el cálculo de EuroSCORE y EuroSCORE II, fueron realizados de forma exclusiva por el autor del estudio.

4.5.3. Fase prospectiva

De forma simultánea a la recogida retrospectiva de datos, se introdujeron de forma prospectiva los registros de los pacientes intervenidos a partir del año 2012. Éstos fueron igualmente extraídos de las historias clínicas e informes preoperatorios. La recopilación de datos de esta fase del estudio, así como el registro de los fallecimientos se realizó mediante seguimiento personal de los pacientes intervenidos por parte del autor del trabajo de forma exclusiva.

4.6. Cálculo de EuroSCORE y EuroSCORE II

Los resultados de EuroSCORE y EuroSCORE II para cada paciente fueron calculados personalmente por el autor del estudio, en base a las variables recogidas desde las diversas fuentes descritas. No se emplearon los valores calculados por otros miembros del Servicio, ni los criterios de otras personas a la hora de interpretar y computar los factores de riesgo. Las definiciones de cada factor de riesgo se siguieron de modo estricto y preciso, tal y como fueron publicadas en los trabajos originales de EuroSCORE y EuroSCORE II (8, 10).

Fueron utilizados exclusivamente los valores de la versión logística del EuroSCORE original, que fueron empleados para el cálculo de la mortalidad ajustada el riesgo y la construcción del gráfico de embudo.

Los valores de las variables de EuroSCORE II fueron utilizados para el análisis descriptivo, y se emplearon para el cálculo de la probabilidad de mortalidad peroperatoria esperada del grupo de estudio, así como para el análisis de la calibración y discriminación posteriores.

Todos los datos fueron sometidos a una prueba de coherencia mediante las hojas de cálculo previamente descritas.

4.7. Análisis estadístico

4.7.1. Estadística descriptiva

Las variables continuas del estudio son el predictor “edad” y el valor medio del resultado de EuroSCORE II, expresados como media +/- desviación estándar y rango. Las variables categóricas son el resto de variables secundarias definidas anteriormente, y se expresaron en forma de porcentaje. La variable resultado “mortalidad peroperatoria” se expresa también en forma de tanto por ciento.

4.7.2. Comparación con datos nacionales. Gráfico de embudo

La comparación de los resultados de mortalidad obtenidos en el estudio se realizó con la misma metodología empleada en el Primer Informe del Proyecto de Español de Calidad de Cirugía Cardiovascular del Adulto (121), definida por Spiegelhalter (31). Para ello se ajustó la mortalidad observada en base al cociente entre mortalidad observada y esperada (O/E). El modelo empleado para el cálculo de la mortalidad esperada fue la versión original de EuroSCORE, al ser el más empleado en España, y haber sido elegido como patrón de referencia en el mayor proyecto en Calidad realizado en nuestro medio, ya que fue considerado como único modelo validado (aunque de forma indirecta) en España hasta el momento, al estar la nueva versión en fase de estudio durante el desarrollo de este proyecto. La mortalidad de referencia empleada fue la mortalidad española media observada en dicho proyecto de calidad. La mortalidad ajustada al riesgo se ajustó en base de *odds* para que los valores se ajustasen al intervalo (0,1) y no fuese posible obtener valores de mortalidad superiores al 100%. Los cálculos se realizaron de la siguiente forma:

- Se dedujo la mortalidad cruda nacional a partir de los datos del Primer Informe del Proyecto de Español de Calidad de Cirugía Cardiovascular del Adulto, que se utilizó como referencia (R).

- Se calculó la mortalidad observada del grupo del Hospital General Universitario de Alicante (O) y la mortalidad esperada, que es la media de los valores de EuroSCORE para los pacientes del grupo del estudio (E). La proporción observado/esperado (O/E) es el cociente entre ambos valores.

- La mortalidad puede ser interpretada como probabilidad de muerte de un paciente. En teoría de probabilidades, una probabilidad también puede ser expresada como *odds* (cociente del número de eventos a favor entre el número de eventos en contra). En base a esto calculamos la *odds* para la mortalidad de referencia (R) de la siguiente forma:

$$odds\ (R) = \frac{R}{(1-R)}$$

- Con estos datos se calculó la mortalidad ajustada al riesgo (MAR) del grupo de estudio mediante la siguiente expresión en base a la *odds*:

$$MAR = \frac{(O/E)\times odds(R)}{1+((O/E)\times odds(R))}$$

- Para evaluar gráficamente la concordancia del valor del grupo de estudio con el del grupo de referencia se utilizó un gráfico de embudo (*funnel plot*), que se construyó representado en el eje de abscisas el número de pacientes del grupo de estudio, y en el eje de ordenadas las distintas mortalidades representadas por tres líneas, una paralela al eje de abscisas que interseccionа con el de ordenadas en el valor de la mortalidad de referencia (R) , y las dos líneas curvas convergentes hacia la línea previa que indican los límites de alarma superior e inferior. Se calculó un intervalo de confianza al 99% respecto de R, siendo el parámetro de precisión del intervalo el tamaño de la muestra. Los límites del intervalo (y_{99}) se calcularon mediante la siguiente expresión:

$$y_{99} = R + (-2{,}58)\sqrt{\frac{R(1-R)}{n}}$$

siendo -2,58 el número de desviaciones estándar necesarias para calcular el intervalo de confianza al 99% y *n* el tamaño de la muestra del grupo de estudio. La raíz cuadrada representada tiene un resultado positivo y otro negativo, empleándose cada uno de ellos para construir cada uno de los intervalos, el superior y el inferior. Aplicando múltiples valores de *n* se lograron un igual número de valores de y_{99} que, representados gráficamente, construyeron las líneas curvas de los límites de intervalo

antes descritos con la forma característica de embudo que da nombre a este tipo de gráfico.

- Por último se representó el valor de la MAR del grupo de estudio dentro de la gráfica de embudo como un punto cuya coordenada en el eje de abscisas corresponde al tamaño de la muestra empleada. La interpretación de este tipo de gráfico es que puede considerarse que no existen diferencias significativas con la mortalidad de referencia si la MAR está comprendida entre las líneas que delimitan los límites del intervalo de confianza.

4.7.3. Calibración de EuroSCORE II

Para realizar la calibración de EuroSCORE II se empleó el estadístico de bondad de ajuste de Hosmer-Lemeshow. Éste fue elegido al ser la prueba empleada en el trabajo original de EuroSCORE II para llevar a cabo la validación interna del modelo (10), así como el más empleado en la literatura en estudios de validación de modelos predictivos (23). Se expresaron los resultados de mortalidad observada y esperada ordenados por deciles de riesgo, el valor del estadístico de contraste y el valor de p obtenido.

4.7.4. Discriminación de EuroSCORE II

Para el cálculo de la capacidad de discriminación de EuroSCORE II se construyó una curva ROC y se obtuvo el valor del área bajo la curva, que se expresó con un intervalo de confianza al 95%. De forma similar, ésta fue la prueba empleada para valorar la discriminación de EuroSCORE II en la validación interna del trabajo original (10).

5. RESULTADOS

Se incluyeron en el estudio 970 pacientes intervenidos en el centro de estudio entre el 1 de enero de 2010 y el 31 de agosto de 2013, respetando los criterios de inclusión y exclusión previamente expuestos. La exclusión de los pacientes que fueron incluidos en el estudio original de EuroSCORE II (los intervenidos entre mayo y julio de 2010) supuso una pérdida de 71 pacientes para el estudio actual.

La recopilación de los datos de los años 2010 y 2011 se realizó de forma retrospectiva, y la de 2012 y 2013 de forma prospectiva. La compleción de los registros fue total, con una tasa de datos perdidos o incoherentes de cero.

A pesar de que el proyecto originalmente contemplaba la recogida de datos hasta el final del año 2013, completando un periodo de inclusión de 4 años, ésta tuvo que ser suspendida el 31 de agosto de 2013 por el cese de la relación laboral del autor del proyecto con el Servicio en que se realizaba el estudio.

5.1. Datos descriptivos

La edad media de los pacientes fue de 66,4 ± 11,3 años (rango 17-87). La proporción de hombres y mujeres fue del 62,4% y 37,6%, respectivamente.

Respecto a las variables de predicción de EuroSCORE II, la prevalencia de enfermedad pulmonar crónica fue del 9,8%, de arteriopatía extracardiaca del 11,6%, de movilidad reducida del 1,8% y de disfunción renal del 63,8% (aclaramiento de creatinina entre 51-85 mL/min, 44,1%; de 50 mL/min o menor, 18,7%; hemodiálisis crónica, 1,0%). Hubo un 6,2% de reintervenciones, un 1,6% de pacientes intervenidos en fase aguda de endocarditis y el 2,2% se intervino en un estado crítico según definición de EuroSCORE II (10). Un 6,5% de pacientes presentaban angina de reposo preoperatoria, y el 18,0% tenían algún grado de disfunción ventricular izquierda (fracción de eyección entre 31 y 50% el 14,6%, entre 21 y 30% el 1,9% y menor del 21% el 1,5%). La proporción de pacientes

intervenidos tras un infarto agudo de miocardio fue del 10,7%. El 34,0% de los pacientes presentaba hipertensión pulmonar (moderada en el 23,4% de los casos y grave en el 10,6%); en cuanto a la prioridad de la intervención, el 33,2% de los pacientes se intervinieron de forma preferente (durante el ingreso hospitalario sin poder ser dado de alta previamente a la intervención por su estado clínico), el 3,6% requirió una intervención urgente y no hubo ningún caso de cirugía de salvamento. En cuanto al tipo de intervención realizada, un 31,8% de los casos fueron intervenciones coronarias aisladas, un 33,7% fueron procedimientos únicos no coronarios, un 25,7% dobles procedimientos y el 8,8% recibieron tres o más procedimientos. Se realizó cirugía sobre la aorta torácica en el 7,0% de los pacientes. El 9,4% de los pacientes presentaban diabetes mellitus dependiente de insulina. Por último, el 32,8% de los pacientes presentaban una clase funcional grado I según la NYHA, mientras que el 26,4% se encontraban en clase II, el 34,2% en clase III y el 6,8% en clase IV.

Estos datos están resumidos en la Tabla 4.

Tabla 4. Análisis descriptivo de factores de riesgo para EuroSCORE II

Variable de predicción	Frecuencia (%) o media (DE, rango)	Variable de predicción	Frecuencia (%) o media (DE, rango)
Edad	66,4 (11,3, 17-87)	IAM	10,7%
Sexo femenino	37,6%	HTP 31-55 mmHg	23,4%
Enfermedad pulmonar	9,8%	HTP > 55 mmHg	10,6%
Arteriopatía extracardiaca	11,6%	Intervención preferente	33,2%
Movilidad reducida	1,8%	Intervención urgente	3,6%
CrCl 51-85 mL/min	44,1%	Intervención salvamento	0%
CrCl <51 mL/min	18,7%	1 no coronario	33,7%
Hemodiálisis	1,0%	2 procedimientos	25,7%
Cirugía cardiaca previa	6,2%	3/+ procedimientos	8,8%
Endocarditis activa	1,6%	Aorta torácica	7,0%
Estado preoperatorio crítico	2,2%	DMID	9,4%
Angina de reposo	6,5%	NYHA II	26,4%
FEVI 31-50%	14,6%	NYHA III	34,2%
FEVI 21-30%	1,9%	NYHA IV	6,8%
FEVI <21%	1,5%		

DE: desviación estándar; CrCl: aclaramiento de creatinina; FEVI: fracción de eyección del ventrículo izquierdo; HTP: hipertensión pulmonar; DMID: diabetes mellitus insulín-dependiente; NYHA: New York Heart Association

5.2. Mortalidad observada

La mortalidad intrahospitalaria observada en el estudio, de acuerdo con la definición del trabajo original de EuroSCORE II (10), fue de 82 pacientes, lo que supone un 8,5% ± 2,8% de los pacientes intervenidos.

5.3. Comparación con datos nacionales. Gráfico de embudo

Para comparar los resultados del centro del estudio con los resultados publicados a nivel nacional se calculó la mortalidad ajustada al riesgo del grupo de estudio, que fue representada dentro de un gráfico de embudo, siguiendo la metodología previamente expuesta (31). El valor de la mortalidad ajustada al riesgo obtenido fue de 7,2% en base a EuroSCORE y los datos de mortalidad nacional recogidos en el Primer Informe del Proyecto de Español de Calidad de Cirugía Cardiovascular del Adulto, donde se publicó un mortalidad del 5,7%.

En este gráfico se aprecia que la mortalidad ajustada al riesgo del Servicio de Cirugía Cardiovascular del Hospital General Universitario de Alicante está dentro de los parámetros de confianza para el tamaño de muestra estudiado, y que por tanto no difiere significativamente de la mortalidad media nacional observada en el estudio de referencia (Figura 5).

5.4. Mortalidad esperada

El valor medio obtenido para EuroSCORE II, que refleja la mortalidad estimada por el modelo para el grupo de estudio fue de 3,6% ± 4,1%, (rango 0,5-34,9).

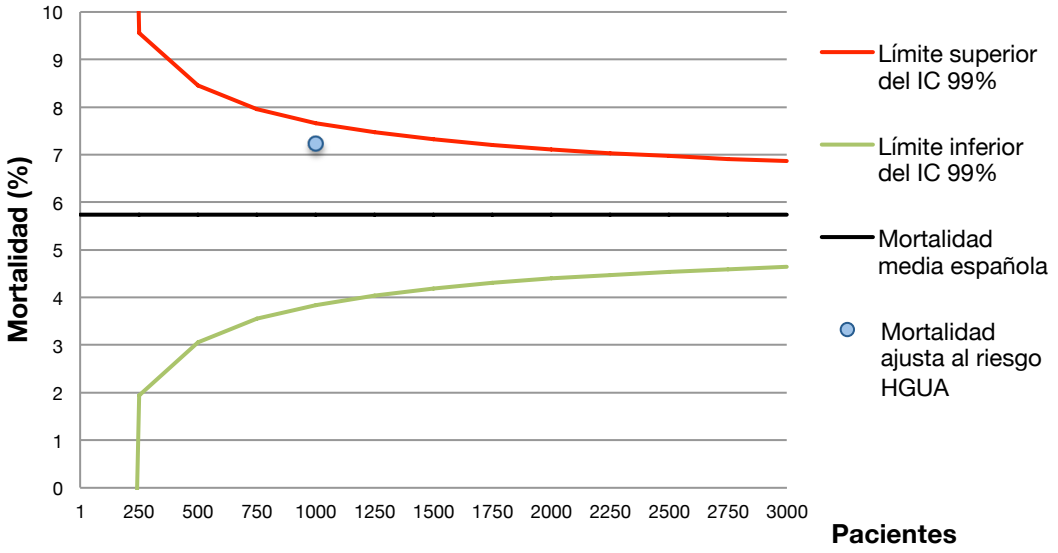


Figura 5. Gráfico de embudo de la mortalidad ajustada al riesgo en el Hospital General Universitario de Alicante

5.5. Calibración de EuroSCORE II

La prueba de bondad de ajuste de Hosmer-Lemeshow proporcionó un valor de 13,6, con un valor de p de 0,09. Este valor implica que no existe una diferencia estadísticamente significativa entre la mortalidad observada y esperada, aunque está cerca de la significación estadística, y por tanto cercano al fallo de calibración por infraestimación del riesgo.

Los valores esperados y observados, clasificados por deciles de riesgo se pueden apreciar en la Tabla 5.

Tabla 5. Distribución de valores por deciles de riesgo en la prueba de bondad de ajuste						
		VIVOS		FALLECIDOS		Total
		Observado	Esperado	Observado	Esperado	
Deciles de riesgo	1	93	89,9	2	5,0	95
	2	95	90,7	1	5,3	96
	3	95	93,3	4	5,7	99
	4	96	93,1	3	5,9	99
	5	91	91,8	7	6,2	98
	6	89	91,4	9	6,6	98
	7	89	89,9	8	7,1	97
	8	85	89,8	13	8,2	98
	9	84	87,8	14	10,2	98
	10	71	70,1	21	21,9	92

Como se aprecia en la tabla, el número de muertes es menor del esperado en deciles de bajo riesgo (sobreestimación), pero es mayor en a partir del quinto decil (infraestimación en segmentos de alto riesgo).

5.6. Discriminación de EuroSCORE II

Para la valoración de la capacidad de discriminación del modelo, se construyó la curva ROC según la metodología previamente indicada (Figura 6).

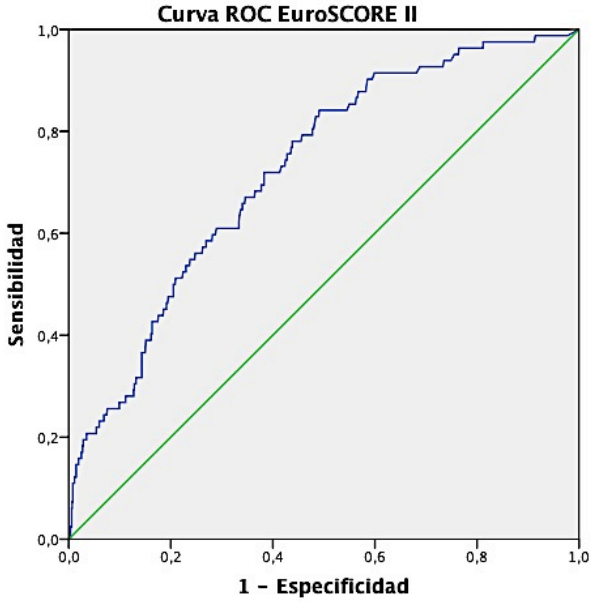


Figura 6. Curva ROC para EuroSCORE II

El valor obtenido para el área bajo la curva (estadístico *c*) fue de 0,72 (intervalo de confiaza al 95%: 0,67-0,78). Este resultado implica una capacidad de discriminación adecuada para el modelo. El intervalo de confianza al 95% se aleja suficientemente de 0,5, valor que implicaría la ausencia de discriminación del modelo, siendo el límite inferior muy próximo a 0,7, valor que se considera adecuado para un modelo, y el superior a 0,8, que indica un muy buen resultado de discriminación.

6. DISCUSIÓN

La pertinencia y necesidad de este estudio viene justificada tanto por los resultados contradictorios de los que se dispone actualmente acerca del rendimiento de EuroSCORE II como por el concepto básico de que cualquier modelo predictivo debe de ser validado de forma externa para evaluar su rendimiento real, especialmente si las características de la población estudiada difieren significativamente de las de la muestra a partir de la cual se desarrolló el mismo.

En estos términos debemos destacar los trabajos realizados por Silva y García-Valentín. El primero, ya comentado con anterioridad (97), es una validación externa, ambispectiva y realizada en un solo centro, en el que EuroSCORE II presentó un fallo de calibración por infraestimación del riesgo. En la misma dirección, y auspiciado por la SECTCV, García-Valentín et al. han publicado recientemente un estudio prospectivo y multicéntrico de validación de EuroSCORE II de ámbito nacional en nuestro país (99), cuyos resultados han sido similares a los de Silva. Una ampliación de este trabajo que evalúa la funcionalidad de EuroSCORE y EuroSCORE II como modelos de referencia de mortalidad ha sido también recientemente publicada (100), generando importante repercusión y comentarios (23, 24, 126, 127). Este último trabajo sugiere que EuroSCORE II es un mejor modelo de referencia en nuestro medio que su primera versión, aunque ha de utilizarse con precaución por la existencia de una tendencia a la infraestimación del riesgo quirúrgico.

Considerando que las características de la población atendida, complejidad de procedimientos y resultados quirúrgicos pueden tener amplias variaciones entre hospitales de medio y alto número de intervenciones, consideramos que la realización de un estudio de estas características está justificada con el objetivo de evaluar el rendimiento real de este modelo en relación al volumen de actividad de las unidades quirúrgicas.

El estudio de validación más amplio realizado hasta el momento en España, y que analizaremos como referencia de los resultados de este trabajo, es el publicado por García-Valentín et al (99, 100). Las conclusiones que se extraen de este estudio son, en primer lugar, el empeoramiento del perfil de riesgo de los pacientes intervenidos en España. El valor medio de la mortalidad esperada por EuroSCORE II fue de 5,7% en ese estudio, sensiblemente superior a la mortalidad esperada para la muestra del proyecto aquí expuesto (3,6%). La diferencia entre ambos valores de mortalidad esperada es notable, y probablemente atribuible a las diferencias en la complejidad de los pacientes entre las dos muestras. Es este sentido, no es desestimable el hecho de que los datos de EuroSCORE II en esta tesis hayan sido calculados por una única persona siguiendo estrictamente los criterios y definiciones del artículo original, mientras que los datos del estudio de validación nacional han sido recogidos por múltiples autores, estando sometidos a una mayor variabilidad de interpretación de los criterios, y que puede haber influido en la obtención de esos resultados.

6.1. Obtención de datos y tamaño de la muestra

El número de eventos resultado recogidos ha sido menor que el estimado en las expectativas iniciales. Este hecho ha generado una pérdida importante de potencia estadística, más relevante en la prueba de calibración.

El descenso en el volumen de actividad quirúrgica del Servicio durante algunos meses del año 2012 por motivos de gestión hospitalaria, así como la imposibilidad de acceso a los datos del último cuatrimestre de 2013 por motivos ya mencionados han sido las principales causas de esta disminución en el número de pacientes reclutados respecto a la previsión inicial.

La tasa de datos perdidos o incoherentes ha sido cero. Este hecho implica una alta calidad de los datos obtenidos para la realización de este estudio, que son acordes con la calidad de los datos obtenidos en otros proyecto realizados en España (99, 100, 121).

6.2. Datos demográficos y factores de riesgo

El análisis comparativo entre los datos demográficos y descriptivos de las variables predictoras de EuroSCORE II entre este estudio y el de validación nacional está recogido en la Tabla 6:

Tabla 6. Comparación de factores de riesgo entre el estudio actual y el de validación nacional de EuroSCORE II

	Validación nacional	Validación medio volumen
Variable	Frecuencia (%) o media (DE)	Frecuencia (%) o media (DE)
Edad	66,6 (12,3)	66,4 (11,3)
Sexo femenino	36,2%	37,6%
Enfermedad pulmonar	8,3%	9,8%
Arteriopatía extracardiaca	11,9%	11,6%
Movilidad reducida	3,7%	1,8%
CrCl 51-85 mL/min	42,2%	44,1%
CrCl <51 mL/min	23,2%	18,7%
Hemodiálisis	0,8%	1,0%
Cirugía cardiaca previa	6,7%	6,2%
Endocarditis activa	3,1%	1,6%
Estado preoperatorio crítico	6,8%	2,2%
Angina de reposo	7,5%	6,5%
FEVI 31-50%	9,8%	14,6%
FEVI 21-30%	3,3%	1,9%
FEVI <21%	7,3%	1,5%
IAM	12,2%	10,7%
HTP moderada	21,6%	23,4%
HTP grave	12,3%	10,6%
Intervención preferente	39,2%	33,2%
Intervención urgente	4,5%	3,6%
Intervención salvamento	0,2%	0%
1 no coronario	39,9%	33,7%
2 procedimientos	28%	25,7%
3/+ procedimientos	6,7%	8,8%
Aorta torácica	9,2%	7,0%
DMID	6,7%	9,4%
NYHA II	35,3%	26,4%
NYHA III	35,7%	34,2%
NYHA IV	8,7%	6,8%

DE: desviación estándar; CrCl: aclaramiento de creatinina; FEVI: fracción de eyección del ventrículo izquierdo; HTP: hipertensión pulmonar; DMID: diabetes mellitus insulín-dependiente; NYHA: New York Heart Association

A pesar de que los datos son muy similares, existen ciertas variables de gran peso dentro del modelo en las que pueden observarse diferencias notables entre ambos estudios. La proporción de pacientes intervenidos en situación de endocarditis activa (bajo tratamiento antibiótico) y en estado preoperatorio crítico (intubados, con soporte vasoactivo o balón de contrapulsación intraaórtico, en fracaso renal, o con masaje cardiaco previo) son perceptiblemente más altas en la muestra empleada para la validación nacional. Además resulta llamativa la alta proporción de pacientes con fracción de eyección del ventrículo izquierdo inferior al 21% en este mismo estudio, variable con un importante peso dentro del modelo. Éste es un dato habitualmente obtenido por medio de diversas exploraciones (ecocardiografía, ventriculografía, resonancia magnética, etc.), sometida por tanto a una amplia variabilidad inter e intra observador, por lo que la explicación más lógica para esta importante diferencia es la heterogeneidad de los datos del estudio de validación nacional frente a los del estudio aquí expuesto. Asimismo se pueden apreciar diferencias no tan marcadas entre ambos grupos en cuanto a pacientes en peor clase funcional de la NYHA, mayor proporción de cirugía de la aorta torácica o de pacientes intervenidos de forma urgente, siempre a favor del aumento de complejidad de la muestra del estudio multicéntrico nacional de validación.

6.3. Mortalidad observada. Gráfico de embudo

La mortalidad peroperatoria observada en este estudio ha sido del 8,5%. Esta cifra es consistente, aunque ligeramente más alta, con los datos existentes en otros estudios más amplios realizados en nuestro país en materia de evaluación de calidad y validaciones de modelos predictivos. En el Primer Proyecto de Calidad de 2013 (121) la mortalidad registrada fue del 5,7%, mientras que tanto en el Proyecto de Validación de EuroSCORE II en España (99, 100) como en la validación de Silva et al. (97) se observaron mortalidades prácticamente idénticas, del 6,5% y 6,3% respectivamente.

El gráfico de embudo ha sido la herramienta empleada para evaluar los resultados de mortalidad del Servicio de Cirugía Cardiaca del Hospital General de Alicante mediante el parámetro de mortalidad ajustada al riesgo y en comparación con los datos de mortalidad nacional del Primer Informe del Proyecto Español de Calidad de Cirugía Cardiovascular del adulto (121).

En la Figura 5, expuesta en los resultados del estudio, se puede observar que la localización del punto que ilustra la mortalidad ajustada al riesgo del centro del estudio está dentro de los márgenes de confianza del valor de referencia. Existe una tendencia a que la mortalidad se sitúe cerca del límite superior de confianza. La interpretación más objetiva de estos datos es que la mortalidad que se desprende de la actividad del Servicio estudiado es razonablemente similar a la declarada en el Primer Informe del Primer Informe del Proyecto Español de Calidad de Cirugía Cardiovascular del adulto y resto de estudios nacionales, con una ligera tendencia no significativa hacia el incremento de mortalidad.

Un posible sesgo en este resultado viene dispuesto por la elección de la mortalidad de referencia, posiblemente infraestimada ya que en el registro empleado la participación fue voluntaria, hecho que normalmente selecciona datos de los centros con mejores resultados quirúrgicos, al ser los Servicios de peores resultados menos proclives a comunicar su mortalidad.

El reciente estudio de Validación de EuroSCORE II en España (99, 100) obtuvo un dato de mortalidad observada similar, aunque ligeramente superior al declarado en dicho Informe (6,5% frente al 5,7%). A pesar de ser otro estudio de participación voluntaria, los datos fueron revisados y analizados hasta obtener un 0% de datos perdidos o incoherentes, por lo que se puede asumir una gran calidad de los mismos. Si utilizamos el dato de mortalidad observada en este proyecto como referencia en la construcción del gráfico de embudo, observamos que el punto de mortalidad ajustada al riesgo del Servicio analizado está en una posición muy similar a la del gráfico previo, cercano al límite superior (Figura 7).

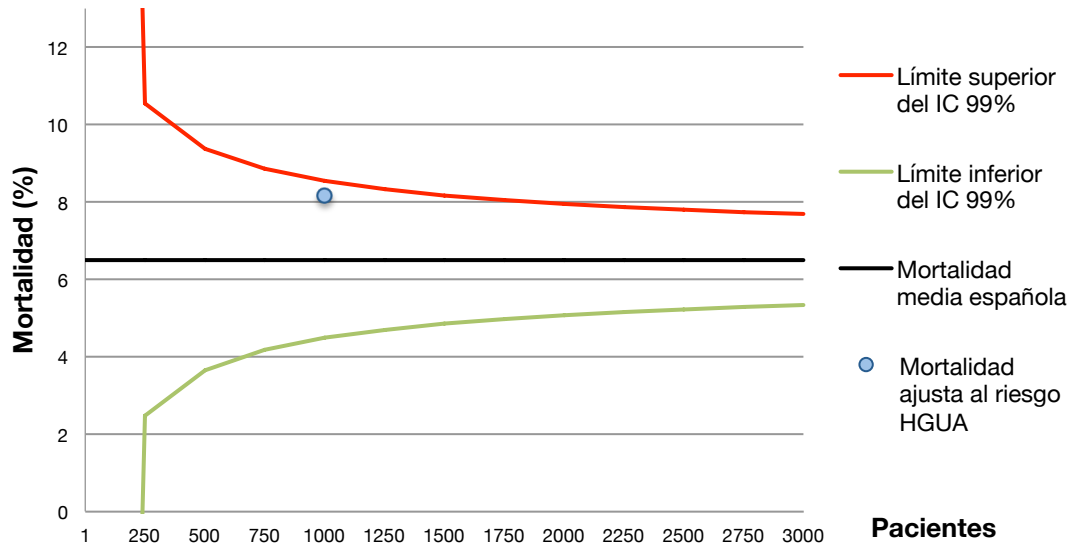


Figura 7. Gráfico de embudo utilizando como referencia el estudio de Validación de EuroSCORE II en España.

Esta tendencia no significativa hacia el aumento de mortalidad puede explicarse por diferentes motivos; el volumen medio de carga quirúrgica del centro puede justificar este hecho, como se menciona en la hipótesis de este trabajo, así como un tamaño de muestra situado en la parte inicial de la gráfica, sujeto a una aleatoriedad significativa. Una muestra más amplia que conservase el mismo valor de MAR situaría este dato por encima del límite de alarma superior, aunque el resultado de mortalidad del periodo de estudio puede haberse visto aumentado por distintas causas, como se comentará más adelante, manteniendo la tendencia global del Servicio valores inferiores de mortalidad.

A pesar de que la metodología seleccionada para este estudio ha sido la misma empleada por los autores del más reciente informe del registro europeo de actividad quirúrgica (128) y el mayor trabajo de calidad previamente realizado en España (121), ésta puede ser puesta en entredicho por los resultados del proyecto de Validación de EuroSCORE II en España, en su ampliación del estudio publicada en la *European Journal of Cardio-Thoracic Surgery* (100). Este trabajo

sugiere que la versión original de EuroSCORE, que es la utilizada como referencia en los estudios mencionados para la construcción de gráficos de embudo, puede contener un problema de sobreestimación del riesgo en la población española que la invalide como herramienta de predicción, y por tanto como referencia válida en la construcción de este tipo de gráficos y en el control de la calidad de la asistencia quirúrgica en general. A pesar de ello estas presunciones derivan de los datos de un solo estudio, y sería necesario realizar más estudios en poblaciones seleccionadas para determinar con precisión la validez de cada modelo predictivo como referente en cada situación específica.

Los resultados del trabajo que se presenta sugieren que la mortalidad ajustada al riesgo del centro del estudio no difiere significativamente de la mortalidad media extraída de los dos estudios multicéntricos de calidad asistencial en Cirugía Cardiovascular más recientemente publicados en nuestro país, utilizando como marco de referencia el modelo de predicción actualmente validado para la metodología descrita.

6.4. Discriminación de EuroSCORE II

El área bajo la curva ROC muestra una capacidad de discriminación adecuada para la población del estudio, con una valor de 0,72 y un intervalo de confianza al 95% entre 0,67 y 0,78. El intervalo de confianza en suficientemente estrecho, alejándose el límite inferior del valor de 0,5, que traduciría la ausencia de discriminación del modelo.

El valor del área bajo la curva ROC es sensiblemente menor que el comunicado por los autores en el trabajo original de EuroSCORE II, que fue de 0,81 (0,78-0,84) (10). En estudios nacionales, la evaluación de la discriminación de este modelo ha obtenido resultados globales satisfactorios. El trabajo de Silva et al. obtuvo un área bajo la curva ROC de 0,78 (0,75-0,81) (97), más cercana al de la validación interna del manuscrito original del modelo y con un intervalo de confianza ligeramente más estrecho al obtenido en este estudio, debido al mayor tamaño de

la muestra. Asimismo el estudio de validación de EuroSCORE II en España obtuvo un área de 0,79 (0,76-0,82) (99, 100), muy similar a la de Silva et al.

Otros estudios de validación realizados fuera de España también han mostrado resultados de discriminación muy similares a los estudio original, como el estudio retrospectivo y multicéntrico de validación en Italia de Barili et al. (91), el estudio argentino de un centro de Borracci et al. (30), de pequeño tamaño y también retrospectivo, el amplio metaanálisis de más de 145.000 pacientes publicado por Guida et al. (129) o el estudio de validación de EuroSCORE II en un centro húngaro (130).

Aunque no está contemplado entre los objetivos del estudio, con los datos obtenidos es posible calcular el área bajo la curva ROC para la versión original de EuroSCORE, que tiene un valor de 0,71 (intervalo de confianza al 95%: 0,65-0,76). El tamaño de la muestra no ha sido estimado para realizar comparaciones entre ambas curvas ROC, cálculo que precisa de un gran número de pacientes. Aun así podemos aproximar de una forma cualitativa y subjetiva que los intervalos de confianza de ambos modelos se superponen ampliamente, lo que sugiere que la discriminación de ambos modelos es muy similar en la población del estudio. Este dato es congruente con los datos de validación interna del estudio original (10), que describió una ligera mejoría no significativa en la capacidad de discriminación respecto a la versión previa. Asimismo estos resultados dan soporte a los datos de estudios de discriminación entre los dos modelos en proyectos de validación ya mencionados (97, 130), a pesar de que la mayoría de ellos dudosamente alcanzan la potencia estadística necesaria para ello debido a tamaños de muestra insuficientes. Curiosamente existen subgrupos en ciertos estudios es los que se obtiene mejor capacidad de discriminación con la antigua versión de EuroSCORE (89), aunque son datos anecdóticos. La versión original del modelo no mostró problemas en su capacidad de discriminación a lo largo de los distintos estudios existentes en la literatura (131), incluso mostrándose superior a otros modelos existentes en la época (132).

A pesar de que los resultados son ligeramente peores en nuestro medio que en otros estudios más amplios, los resultados de este trabajo muestran que la capacidad de discriminación del nuevo modelo es adecuada para centros de medio volumen, aunque más limitada que en otros entornos.

6.5. Calibración de EuroSCORE II

Los resultados obtenidos en la evaluación de la calibración de EuroSCORE II en la muestra del estudio requiere un análisis juicioso. Es evidente que, independientemente del valor de p y del resultado del estadístico de contraste, existe una diferencia clínicamente relevante entre la mortalidad observada (8,5%) y la esperada por EuroSCORE II (3,6%).

La prueba de bondad de ajuste de Hosmer-Lemeshow muestra un valor de 13,6 con un valor de p de 0,09. Este valor se encuentra por encima del límite establecido para determinar la significación estadística (o error α), que es de 0,05, por lo que podríamos afirmar que no existen diferencias estadísticamente significativas entre la mortalidad esperada y observada.

A pesar de ello resulta obvio que sí existe una diferencia notoria entre ambos valores, de forma que la mortalidad observada supera en más del doble a la esperada. La obtención de un resultado de la prueba de calibración como el obtenido, a pesar de las diferencias existentes, puede ser explicada por diferentes motivos, siendo el fundamental el insuficiente tamaño de la muestra del estudio; la estimación del mismo para estudios de validación puede realizarse de forma precisa para el cálculo del área bajo la curva ROC, según la metodología expuesta en los trabajos de Hanley y McNeil (27), siendo su cálculo más complejo para la prueba de calibración. A pesar de que la prueba de Hosmer-Lemeshow ha sido la más empleada en los distintos trabajos de validación a través de la literatura existente, a día de hoy se encuentra en entredicho por el hecho de ser muy sensible al tamaño de la muestra del estudio. Esta característica es común a todas las pruebas estadísticas derivadas del χ^2, como ocurre en este caso (22). De

hecho el mismo Lemeshow firma junto a otros autores un trabajo de reciente publicación en el que demuestra esta dependencia, además de instruir acerca de las características idóneas del tamaño de la muestra y número óptimo de grupos de riesgo empleados en el análisis que se debe utilizar. La aplicación de estas recomendaciones a los cálculos de este estudio nos indica, en primer lugar, que se ha escogido el número de grupos correcto (10 grupos de riesgo, que son los que se emplean en el cálculo por defecto en la mayoría de aplicaciones informáticas existentes), y en segundo lugar que una muestra cercana a 1.000 pacientes con la mortalidad obtenida nos otorga una potencia estadística relativamente pequeña, de entre el 30-40%. Son necesarias muestra del doble de pacientes (aproximadamente 2.000) para obtener una potencia estadística del 70%-80%, aunque esto es dependiente de la frecuencia estimada de aparición del resultado (25). En estos términos es comúnmente aceptado que un estudio de calibración debe contener al menos 100 eventos para obtener una potencia estadística adecuada (3). En el periodo de estudio se han podido registrar 82 muertes postoperatorias, cifra cercana a la recomendada, aunque insuficiente a la luz de los resultados obtenidos.

El valor de p obtenido es muy cercano al valor de significación estadística, y podemos asumir que si se hubiese obtenido un número de eventos resultado tan solo ligeramente mayor, habríamos encontrado diferencias estadísticamente significativas entre la mortalidad observada y la esperada. Por otro lado existen opiniones que defienden que el valor de p en la prueba de Hosmer-Lemeshow no ha de ser interpretado como un concepto de todo o nada (hay o no diferencias estadísticamente significativas), sino que un valor de p muy cercano al valor de significación, como en el caso de este trabajo, debe hacernos suponer que existen deficiencias de calibración del modelo en esa muestra. Por el contrario, cuanto más cercano sea el valor de p a uno, mejor calibrado estará el modelo para esa población (2). Desde hace años, existe la opinión de que la calidad de los resultados científicos no debe limitarse al mero hecho de hallar un valor de p lo más bajo posible, sino que ha de considerarse la magnitud y relevancia de las diferencias observadas (concepto denominado *tamaño* del efecto) (133).

Se puede afirmar, por tanto, que con los datos de este estudio, la mortalidad esperada según el modelo de predicción EuroSCORE II muestra una diferencia importante con la mortalidad observada en el centro de estudio, que se encuentra muy cercana a la significación estadística, y que muy probablemente demostraría estadísticamente un fallo de calibración por infraestimación del riesgo quirúrgico si la muestra fuese lo suficientemente amplia como para registrar 100 eventos de fallecimiento postoperatorio.

Los resultados de calibración de EuroSCORE II obtenidos en los mayores estudios de validación realizados en España muestran datos congruentes con los obtenidos en este trabajo. El trabajo de Silva et al. (97) comunicó una mortalidad esperada (3,5%) muy similar a la de este estudio (3,6%), demostrando matemáticamente un fallo de calibración con una muestra de aproximadamente 4.000 pacientes. Sin embargo la mortalidad esperada por el EuroSCORE clásico fue sensiblemente superior a la obtenida aquí (9,1% vs. 6,5%). Curiosamente, en el proyecto de Validación de EuroSCORE II en España (99) se obtuvieron valores tanto de EuroSCORE como de EuroSCORE II sensiblemente más altos que en este trabajo, siendo el de EuroSCORE de 9,8%, muy similar al de Silva. Los resultados de los estudios de validación realizados en otros países muestran resultados muy variables, con mortalidades esperadas por EuroSCORE II más bajas, como el 1,7% de la validación en un centro turco (86), cercanas a los valores aquí obtenidos, como el 3,1% la validación en una base de datos de pacientes norteamericanos realizada por Osnabrugge et al (89), o incluso superiores, como el obtenido en la validación en una región italiana, publicada por Paparella et al (88), en el que se comunicaba un 4,4% de mortalidad estimada por EuroSCORE II. Hay que tener en cuenta la presencia de un posible sesgo de publicación en los resultados que hallamos en la literatura, ya que estudios que comunican menor mortalidad, diferencias significativas o resultados que muestran la excelencia de los centros remitentes, son más frecuentemente publicados que aquellos que no cumplen estas condiciones (18-20).

El análisis de la mortalidad por grupos de riesgo muestra, como se aprecia en la Tabla 5, una tendencia a la sobreestimación de la mortalidad en los deciles de bajo riesgo, con una mortalidad observada por debajo de la esperada, así como una infraestimación en los de alto riesgo, con mortalidades observadas por encima de las esperadas. Este fenómeno también se aprecia en el proyecto de Validación de EuroSCORE II en España, así como en el previamente mencionado estudio de Paparella et al., realizado en Italia. Sin embargo, en el estudio de Grant et al. (134), que mostró una adecuada calibración de EuroSCORE II en población inglesa, con mortalidades estimada y observada muy cercanas (3,4% y 3,1%, respectivamente), se produjo sobreestimación en ambos extremos de los grupos, tanto en los de bajo como en los de alto riesgo quirúrgico.

A pesar de que se podría proponer la realización de análisis por subgrupos de patología o análisis de mortalidad por periodos de tiempo o cirujanos individuales, el tamaño de la muestra en este estudio es insuficiente para la realización de este tipo de cálculos, ya que supondrían una reducción inadmisible de la potencia estadística al precisar una corrección del error α por la realización de comparaciones múltiples (135). El fallo de predicción de EuroSCORE en pacientes de alto riesgo es conocido y ya fue comentado anteriormente. En nuestro estudio se ha apreciado una tendencia del nuevo modelo a la infraestimación del riesgo en estos pacientes, derivada de los datos de la tabla 5, que muestra análisis de grupos de la prueba de calibración. Esto sugiere que tampoco EuroSCORE II parece ser una escala apropiada para la evaluación del riesgo de mortalidad en este subgrupo de pacientes. Estos resultados han sido confirmados en un amplio estudio realizado en población británica, en que incluso se apreció mejor rendimiento de la primera versión del modelo (136), así como recientemente en otro trabajo centrado pacientes sometidos cirugía valvular aórtica (137).

Por tanto, se confirma que la mortalidad esperada según EuroSCORE II en este estudio es concordante con los mismos datos obtenidos en estudios previos, tanto nacionales como de otros ámbitos geográficos, a pesar de la heterogeneidad de resultados que se pueden hallar en la literatura.

6.6. Relación volumen-resultado

Retomando la hipótesis de partida de este trabajo, en el que se asume que existe un relación entre el volumen de actividad de una unidad de Cirugía Cardiovascular y los resultados de calidad del mismo, este estudio muestra unos resultados relevantes.

El primero es que la mortalidad observada en el centro de estudio, catalogado como de medio volumen (aproximadamente 300-350 intervenciones de cirugía cardiaca mayor anuales) no muestra diferencias significativas con la mortalidad comunicada en otras series nacionales, y se encuentra dentro del margen de seguridad de las mismas, aunque con una tendencia al aumento de mortalidad peroperatoria. Este dato soporta la hipótesis del estudio, ya que cabría encontrar esta pequeña diferencia, estadísticamente significativa o no, hacia el exceso de mortalidad en los centros de menor volumen, como así ha sido.

El segundo es que, a pesar de no demostrarse matemáticamente un fallo de calibración, con valores de p cercanos al mismo, existe una notable diferencia entre la mortalidad observada y la esperada por EuroSCORE II. Por tanto podemos sugerir que este modelo de predicción no muestra un buen funcionamiento en centros de estas características. Como se proponía en la introducción e hipótesis de este trabajo, la metodología de desarrollo de EuroSCORE II (comunicación de datos de predictores y mortalidad durante un periodo concreto de tiempo) favorece la predominancia de datos provenientes de centros de alto volumen, otorgando a los datos de centros de medio y bajo volumen una menor representatividad en el modelo.

Por tanto, estos resultados sugieren que en nuestro medio también existe una relación volumen-resultado en cuanto a la actividad de las unidades de Cirugía Cardiovascular, con una tendencia no significativa hacia el aumento de mortalidad en unidades de media actividad, comparado con las medias de las series y registros existentes. Probablemente EuroSCORE II no es un modelo adecuado

para este tipo de Servicios, dado que los cálculos en su desarrollo están más influenciados por las características de los centros de alto volumen.

Esta relación es aplicable además para centros y para cirujanos de forma individual, aunque probablemente por razones diferentes. Un centro de bajo volumen conllevará menor experiencia quirúrgica por parte de los cirujanos del Servicio, aunque también menor implicación del resto de las unidades que contribuyen al manejo de los pacientes sometidos a cirugía cardiaca (unidades de cuidados críticos, cardiología, nefrología, hematología, rehabilitación, etc.) Estos hechos pueden contribuir de forma global a que los resultados quirúrgicos sean peores. Con unas premisas similares, la situación del cirujano de bajo volumen incide de forma fundamental en una menor familiaridad con el manejo de pacientes y un menor grado de experiencia quirúrgica.

La complejidad en la caracterización de la relación entre volumen y resultado puede aumentar por el hecho de que un centro de volumen bajo pueda no tener cirujanos de bajo volumen, sino todo lo contrario, un equipo quirúrgico sometido a una relación intervención/cirujano muy alta, incluso presentando una actividad quirúrgica muy superior a la de los profesionales de los centros de alto volumen, posiblemente con plantillas más hipertrofiadas. Asimismo puede darse la situación de que en centros de bajo o medio volumen, el tratamiento del paciente quirúrgico suponga un tanto por ciento de actividad relativamente más importante para el resto de servicios relacionados con los cuidados postquirúrgicos, comparados con centros de gran volumen, y por tanto los resultados postoperatorios no se vean tan afectados por el volumen global, o incluso sean muy buenos.

Una política que quizá podría contribuir a la mejoría de los resultados en los centros con Cirugía Cardiovascular podría ser la de reducir el número de cirujanos en plantilla a fin de incrementar el número de intervenciones por profesional, y la de crear unidades específicas para el diagnóstico y cuidados postoperatorios de los pacientes sometidos a cirugía cardiaca, con el fin de aumentar el grado de

implicación y experiencia de sus integrantes en el manejo de este tipo de pacientes. Por desgracia la cada vez mayor subespecialización de los cirujanos dentro de disciplinas concretas en el ámbito de la Cirugía Cardiovascular llevan a un incremento de los tamaños de las unidades quirúrgicas, lo que va en contra de este concepto.

6.7. Causas del fallo de calibración por infraestimación

El fallo de calibración de un modelo de predicción por infraestimación del riesgo de mortalidad peroperatoria tiene especiales implicaciones.

Este resultado implica que la mortalidad observada es mayor que la esperada, hecho que puede venir derivado por un aumento real de la mortalidad, un defecto en el desarrollo del modelo predictivo, o una mezcla de ambas situaciones. En este estudio nos encontramos con alta probabilidad en el tercer supuesto, ya que hemos asumido una tendencia hacia el aumento de mortalidad en centros de menor volumen junto con una muy probable escasa representación de los datos de estos centros en el modelo predictivo, lo que puede hacer dudar de su validez externa en esta población.

Las características de la población de pacientes sometidos a cirugía cardiaca en España ha sido diferente a la de otros países de forma constante a los largo de los estudios realizados hasta el momento, presentando en la mayoría de las ocasiones un perfil de mayor riesgo quirúrgico, así como una mayor mortalidad observada. Por otro lado también se ha observado un empeoramiento en el perfil de riesgo de los pacientes sometidos a cirugía cardiaca en nuestro país durante los últimos años (43, 97, 99, 100). Este incremento del riesgo se explica por las características propias del sistema español de provisión de servicios sanitarios; el acceso de los pacientes a la atención especializada, los protocolos cardiológicos y de derivación a tratamiento quirúrgico, etc. Todos estos factores conllevan un mayor riesgo quirúrgico de los pacientes intervenidos en España y son incluso más patentes en hospitales de menor volumen quirúrgico, en la que el acceso a

diagnóstico y tratamiento es todavía más complicado y lento que en otros centros de mayor actividad.

Otra posible causa de un fallo de calibración, no siempre admitida, es un diseño o ejecución inadecuados del proyecto de validación. Un estudio de estas características debe tener en primer lugar un tamaño muestra adecuado para lograr una potencia estadística e intervalos de confianza suficientes, así como utilizar los estadísticos de contraste adecuados para realizar las comparaciones entre grupos. Probablemente tamaños de la muestra demasiado grandes tampoco son adecuados cuando se utilizan las pruebas estadísticas comentadas en este estudio, especialmente la de calibración (prueba de bondad de ajuste de Hosmer-Lemeshow), ya que pueden concluir que cualquier pequeña diferencia observada es estadísticamente significativa. Como se ha mencionado, este hecho puede ser corregido con una adecuada selección de los grados de libertad y grupos de riesgo (25). En la actualidad se propone la utilización de otras herramientas estadísticas consideradas de mayor precisión para la evaluación de la modelos predictivos, como la pendiente de calibración o los índices de reclasificación (*net reclassification improvement*) y de discriminación (*integrated discrimination improvement*) (21, 24).

Otro factor relacionado con un fallo de calibración de una forma importante es el de la adecuada interpretación de los factores de riesgo del modelo de predicción y el correcto empleo del mismo, teniendo este hecho una importancia capital en el cálculo adecuado de la mortalidad esperada. Como ya se comentó inicialmente, la definición de las variables debe ser precisa y exacta, y la aplicación de los criterios del modelo ha de hacerse de forma rigurosa y minuciosa (2), como se he realizado en este estudio.

Resumiendo lo anteriormente expuesto, este estudio tiene un tamaño de muestra adecuado para el cálculo de la capacidad de discriminación, aunque pequeño para la prueba de calibración. La definición de variables ha sido exactamente la indicada en los trabajo originales de los modelos de predicción (8, 10) y los

cálculos e interpretación de los factores de riesgo, así como el resultado de la aplicación de cada modelo en todos los pacientes han sido realizados únicamente por el autor del estudio, evitando la variabilidad entre observadores. Este hecho dota al estudio de un nivel muy importante de coherencia y escasos problemas de interpretación, por lo que consideramos que presenta un diseño adecuado para su propósito. Los estudios con los que se han comparado los datos aquí obtenidos tienen un diseño multicéntrico, por lo que están sujetos a la interpretación de múltiples profesionales. Del mismo modo, en los diferentes estudios publicados es posible la existencia de un sesgo por sobreestimación del riesgo, dada la tendencia a considerar mayor la prevalencia de factores de riesgo inherente a cualquier estrategia de control de resultados (37), lo que puede contribuir a la obtención de valores de mortalidad ajustada al riesgo más favorecedoras para el centro que exporta los datos.

Como último punto de influencia en la obtención de un virtual fallo de calibración por infraestimación del riesgo quirúrgico, debemos comentar la calidad de los datos obtenidos, tanto en este estudio como en el conjunto de datos del grupo de desarrollo de EuroSCORE II. Los datos contenidos en este trabajo, además de las características antes mencionadas acerca de su integridad y rigurosidad en la aplicación de las definiciones exactas de cada factor predictor de mortalidad, han sido analizados y comprobados en busca de incoherencias. El resultado tras el paso de estos filtros ha sido la obtención de un conjunto de registros depurado con cero datos perdidos o incoherentes. La base de datos de EuroSCORE II, sin embargo, fue un proyecto a gran escala con participación de decenas de centros y con una importante cantidad de datos incompletos, especialmente en cuanto al seguimiento de la variable resultado (mortalidad postoperatoria). Este hecho no es subsanable desde el punto de vista matemático, y va en contra de la fiabilidad del modelo. La decisión de cambiar el horizonte de predicción de mortalidad (reducida al tiempo de ingreso postoperatorio) vino definida por esta pérdida de datos, creando un modelo probablemente incompleto y con un reconocida tendencia a la infraestimación del riesgo quirúrgico. Las deficiencias del modelo predictivo suponen otra de las posibles explicaciones de un fallo de calibración,

pudiendo también justificar la variabilidad en los resultados que se han obtenido en el resto de validaciones existentes hasta la fecha.

6.8. Implicaciones clínicas

La finalidad última de la existencia de modelos de predicción en Cirugía Cardiovascular es la de aportar información relativa a las probabilidades de complicaciones graves o muerte con el doble propósito de aportar al cirujano datos útiles para la toma de decisiones en la práctica clínica y de informar de forma adecuada al paciente. De forma adicional, tienen un gran valor en el cálculo de la mortalidad ajustada al riesgo para su utilización en materia de control de calidad asistencial.

Un modelo predictivo de proyección internacional debe disponer de unos datos de validación interna que muestren un adecuado rendimiento, así como posteriormente debe demostrar su validez externa en diferentes poblaciones. En el caso de EuroSCORE II ya hemos comentado que los datos de calibración interna rozaban el fallo, además de la existencia de una tendencia a la infraestimación de la mortalidad, reconocida por los mismos autores (10). Asimismo disponemos de una excesiva diversidad de resultados en estudios de validación externa, que fomenta dudas acerca del adecuado funcionamiento de este modelo en cualquier medio.

Los resultados del estudio de la validación del modelo en España han contribuido a perpetuar o incluso acrecentar este escenario de incertidumbre, hasta el punto de ser publicada junto a un editorial de Silva en el que muestra sus dudas acerca del mejor modelo de predicción en nuestro país (138), proponiendo incluso la realización de validaciones externas en cada centro de forma individual, como se ha realizado en este trabajo, con el fin de evaluar cuál de las dos versiones del modelo funciona mejor en cada unidad concreta. Desde otro punto de vista más amplio, un fallo de validación en una población concreta es una buena oportunidad para realizar una revisión del modelo en base a características

específicas del dominio en que se quiere emplear, mediante el reajuste del intercepto y el coeficiente β de los predictores de acuerdo con el peso específico de cada variable en esa muestra (139).

Los resultados de este proyecto de validación en un centro de medio volumen muestran que EuroSCORE II probablemente no es un buen modelo de predicción en centros de estas características, mostrando datos de discriminación adecuada con unas diferencias muy evidentes en cuanto a mortalidad observada y esperada, que en este estudio no han resultado significativas a causa de una baja potencia estadística. Como se desprende de los datos de mortalidad ajustada al riesgo, en que los resultados de este centro no han diferido sustancialmente de la mortalidad esperada según los patrones actuales, probablemente la versión original del modelo todavía tenga validez en este medio, a pesar de las críticas recibidas y de otros resultados obtenidos en estudios que comprendían un espectro mayor y más heterogéneo de unidades quirúrgicas.

En consonancia con este hecho, los mismos autores de EuroSCORE II comentan en el artículo original la necesidad de adaptar el modelo al funcionamiento y mortalidad de cada unidad concreta mediante al cociente entre mortalidad observada y esperada, que denominan cociente de mortalidad ajustada al riesgo (*risk-adjusted mortality ratio*). Con esta práctica se asume que el funcionamiento de un Servicio es igual, mejor o peor que el definido por el patrón de referencia que establece EuroSCORE II, siendo el cociente de mortalidad ajustada al riesgo el factor que modifica el cálculo de la mortalidad esperada en la evaluación de cada caso concreto, o en la implementación de nuevas técnicas quirúrgicas en pacientes de alto riesgo. El problema fundamental de este razonamiento viene definido por la posible existencia de deficiencias en el modelo de referencia, ya comentadas y refrendadas por los resultados de este proyecto, que establecería un patrón erróneo con el que realizar estas comparaciones.

Los resultados de este estudio nos aportan información acerca del rendimiento de una unidad quirúrgica en cuanto a su actividad global. La extrapolación de los

mismos a subgrupos concretos como la cirugía valvular, coronaria, pacientes con perfil de alto riesgo, etc., no es estrictamente correcta ni acorde a las hipótesis y objetivos de este estudio, por lo que no debe ser realizada.

6.9. Calidad de los datos del estudio

Los datos recopilados para la realización de este proyecto han sido sometidos por el autor a un proceso exhaustivo de depuración y escrutinio, de forma que el resultado final es la ausencia de pérdida de datos o de datos incoherentes.

Este trabajo, en consonancia con otros estudios recientes realizados en España, como el Primer Informe del Proyecto Español de Calidad de Cirugía Cardiovascular del Adulto (121) o el proyecto de Validación de EuroSCORE II en España (99, 100), refrenda la realidad de que la producción de datos científicos en nuestro país es de gran calidad, lejana a la de otros registros internaciones (128) en los que abunda la existencia de datos perdidos o incompletos.

6.10. Limitaciones del estudio

La limitación más importante de este trabajo es el tamaño de la muestra. Los cálculos iniciales de tamaño de la muestra se basaron en el cálculo del área bajo la curva ROC, según los trabajos de Hanley y McNeil (27, 28), teniendo como objetivo obtener un intervalo de confianza lo más estrecho y preciso posible en torno a este valor. Pero este cálculo afecta únicamente a la evaluación de la habilidad discriminatoria del modelo. El cálculo del tamaño de la muestra para la prueba de calibración, sin embargo es más complejo y no queda completamente definido. El ya mencionado artículo en el que participa Lemeshow, con recomendaciones para optimizar el tamaño muestra de su prueba en base a parámetros de potencia estadística (25) fue publicado más de 30 años después del original (22), en una época de desprestigio creciente para esta prueba estadística dentro del mundo científico. Según los parámetros del trabajo mencionado, el número de eventos registrado en este trabajo resulta ligeramente

insuficiente para obtener una potencia estadística (probabilidad de aceptar la hipótesis alternativa cuando ésta es cierta, es decir, de obtener diferencias estadísticamente significativas cuando realmente existen) adecuada. En los trabajos científicos con contraste de hipótesis se considera aceptable una potencia estadística del 80%, que supone un error β (probabilidad de rechazar la hipótesis alternativa cuando es cierta, o lo que el lo mismo, de no obtener diferencias cuando realmente existen) del 20% (140). En este estudio se obtuvieron 84 eventos resultados (muertes postoperatorias), cifra cercana al valor de 100 recomendado en algunos estudios (3), aunque no suficiente, ya que las diferencias entre los valores de mortalidad esperada y observada son evidentes, pero no estadísticamente significativas, con un valor de p de 0,09. Según las tablas de tamaño de muestra del artículo de Paul y Lemeshow, con el número de pacientes y probabilidad de eventos obtenidos en este trabajo, la potencia estadística decrece hasta un 40%, con un error β más alto. La interpretación de este resultado es que la diferencia, por tanto, existe, aunque el tamaño de la muestra es insuficiente para consolidarla estadísticamente, dada la probabilidad de evento registrada. Realizando un cálculo aproximado con la probabilidad de muerte postoperatoria en este centro, el tamaño de la muestra debería incrementarse de los 970 pacientes registrados a aproximadamente 1.175 para obtener 100 eventos de mortalidad. Esta cifra es más cercana al volumen real de pacientes intervenidos por el Servicio en el periodo de estudio (1.041 intervenciones), que habría aumentado aproximadamente a 1.150 si hubiese sido posible completar los datos del año 2013. La decisión de excluir a los pacientes ya incluidos en la validación interna del original, así como las circunstancias particulares del autor y el Servicio en la época del estudio (que se comentará más adelante) impidieron reclutar un número mayor de pacientes.

La segunda limitación importante de este proyecto viene derivada de la recogida retrospectiva de datos para los años 2010 y 2011. A pesar de que todos estos datos fueron recopilados en su totalidad por el autor del estudio de una forma exhaustiva, comprobando todos los registros de cada paciente para garantizar la veracidad y exactitud de los mismos, la recogida retrospectiva de datos conlleva

de forma inherente una mayor probabilidad de errores y pérdida de información (2), principalmente en la prevalencia de factores de riesgo. La pérdida de información en la variable resultado es mucho menos probable debido a la confirmación de ese dato a través de varios registros personales, locales e institucionales. La teórica pérdida de datos podría ocasionar principalmente una infraestimación de la mortalidad esperada en la muestra.

Otra limitación dentro del diseño de este proyecto está causada la variabilidad del volumen quirúrgico dentro del mismo Servicio durante el periodo de estudio. Debido a las condiciones socio-económicas que afectaron al país y la comunidad autónoma durante la realización de este trabajo, la actividad quirúrgica se vio reducida de siete a cinco intervenciones semanales durante la mayor parte del año 2012. Dado que uno el planteamiento principal del estudio se basa en la relación entre el volumen de intervenciones y los resultados quirúrgicos, este hecho pudo haber afectado con un peor rendimiento del Servicio durante el periodo de menor actividad. En efecto, analizando las cifras anuales de mortalidad se aprecia un significativo repunte de la mortalidad durante el periodo de menor actividad comparado con la del resto de años (11% en 2012 vs. 7,6% en el resto de años), no completamente justificado por un aumento similar de la mortalidad esperada (7,2% en 2012 vs. 6,4 % en el resto de años, según EuroSCORE I; 3,9% en 2012 vs. 3,5% en el resto de años según EuroSCORE II), siendo los datos de mortalidad observada y esperada de 2010, 2011 y 2013 más homogéneos que los de 2012. En caso de igualarse la mortalidad de 2012 a la del resto de años del periodo de estudio, el perfil de rendimiento de la unidad se habría ajustado todavía más al valor de mortalidad de referencia utilizado en el gráfico de embudo, y sin riesgo de interpretar que la calidad asistencial no sería adecuada con un tamaño de muestra más grande. Asimismo la prueba de calibración habría mostrado un valor de p más alto, precisándose incluso un tamaño de muestra superior para obtener una potencia estadística adecuada, al disminuir la probabilidad de muerte y el número de eventos. Estos datos, sin embargo, deben ser interpretados con cautela, dado que los tamaños de muestra

de cada año de actividad son muy reducidos, y no se han realizado análisis estadísticos que respalden matemáticamente estas afirmaciones.

6.11. Aspectos futuros

Ante el escenario de confusión en que nos encontramos actualmente, con datos difíciles de interpretar o incluso contradictorios, la mejor opción para dirimir la situación concreta que plantea este trabajo sería la realización de un estudio multicéntrico en unidades de medio volumen para corroborar los resultados de este estudio mediante la obtención de un tamaño de la muestra que proporcionase una mayor potencia estadística sin tener que ampliar de manera excesiva el periodo de inclusión y obtener así un perfil homogéneo de los pacientes a lo largo del tiempo.

En la misma dirección del estudio propuesto, la realización de un trabajo para evaluar la utilidad de las dos versiones de EuroSCORE como modelo de referencia en el control de calidad en centros de medio volumen mediante gráficos de embudo puede resultar de interés, permitiéndonos esclarecer la actual vigencia de la antigua versión del modelo y la idoneidad de la nueva.

En un contexto más amplio, actualmente se han solicitado los datos al equipo de EuroSCORE II para la realización de una comparación estadística de las prevalencias de las distintas variables predictores entre la población de desarrollo del modelo y la del proyecto de validación nacional, que se desarrollará en los próximos meses. Este trabajo nos permitirá averiguar si las variables están distribuidas de una manera similar y las diferencias del peso que tienen para las distintas variables en cada población.

7. CONCLUSIONES

- La discriminación del modelo EuroSCORE II es adecuada en un centro de medio volumen. La prueba de bondad de ajuste muestra una diferencia relevante, aunque este estudio no puede aceptar estadísticamente la hipótesis alternativa de fallo de calibración, al no existir diferencias estadísticamente significativas por una potencia estadística inadecuada.
- La mortalidad quirúrgica en el Servicio de Cirugía Cardiaca del Hospital General Universitario de Alicante no presenta diferencias significativas con la media española.
- La distribución y frecuencias de los factores de riesgo de la muestra del estudio son muy similares a las del proyecto de Validación de EuroSCORE II en España, aunque con algunas diferencias que configuran un perfil de mayor riesgo en ese último trabajo.

8. RESUMEN

8.1. Español

Introducción

La evaluación del riesgo es una parte fundamental en el proceso de toma de decisiones en los pacientes en que se indica un tratamiento quirúrgico. Los modelos de predicción son herramientas globalmente utilizadas que nos ayudan a facilitar este proceso, asimilando los datos del paciente y el procedimiento a realizar para estimar la probabilidad de fallecimiento. Estos pueden ser desarrollados mediante datos prospectivos o retrospectivos, utilizando bases de datos clínicas o administrativas, siendo el mejor diseño metodológico la recolección prospectiva de datos clínicos.

El aspecto más importante en la creación de modelos predictivos en la adecuada selección de variables. Los predictores deben tener una alta prevalencia y asociación estadística con la variable resultado, y su definición ser rigurosa. Los resultados deben ser relevantes y bien definidos, denominándose resultados *duros*, siendo el más utilizado la mortalidad. La pérdida de datos es frecuente en los predictores, aunque pueden ser subsanados mediante técnicas de imputación. Sin embargo, la pérdida de datos de la variable resultado es un problema grave que difícilmente puede ser reparado.

Los modelos de predicción se construyen mediante técnicas de regresión estadística. La más frecuente es la regresión logística binaria, utilizada cuando el resultado es una variable discontinua dicotómica. Existen diferentes tipos de predictores (continuos o no, dicotómicos, politómicos) que se manejan de forma diferente, aunque la práctica más frecuente es transformar el mayor número posible de predictores en respuestas dicotómicas (sí o no).

Existen diferentes problemas específicos del diseño de modelos predictivos, que requieren una observación cuidadosa. El sobreajuste se genera cuando el modelo rinde bien en el conjunto del desarrollo, pero no en la población general. La diferencia entre ambos valores de validación se denomina optimismo, y se observa frecuentemente cuando se requieren múltiples análisis de predictores. El diseño de modelos también puede verse afectado por sesgos de selección, observación o publicación. El último concepto importante a considerar en el desarrollo de modelos predictivos es el de parsimonia, que defiende la idea de construir el modelo funcional más simple posible.

La validación externa de los modelos predictivos es un hecho necesario. La metodología más comúnmente empleada, entre diversas opciones, es la valoración por separado de la calibración (que evalúa si la mortalidad esperada media es igual a la observada) y de la discriminación (que evalúa que los pacientes fallecidos son aquellos de mayor riesgo estimado). Aunque actualmente debatida, la prueba más utilizada para la calibración de modelos es la de bondad de ajuste de Hosmer-Lemeshow, siendo el área bajo la curva ROC (*Receiver Operating Characteristics*) la más común para la discriminación. Un aspecto clave es el cálculo de un tamaño de muestra óptimo, que es especialmente complejo para la prueba de calibración. Se considera adecuado un número de 100 eventos para este propósito. Los fallos de calibración pueden ser explicados de diferentes formas; defectos del modelo, malos resultados reales o errores en el estudio de validación. Estas situaciones requieren un análisis juicioso.

Los gráficos de embudo con intervalos de confianza al 99% son herramientas gráficas estadísticas que se emplean en los estudios de calidad. En ellos se expresa una mortalidad de referencia como una línea horizontal que cruza el eje de ordenadas y dos líneas de alarma convergentes hacia ella en la medida en que crece el tamaño de la muestra (otorgándole su característica forma de embudo). La mortalidad ajustada al riesgo se representa como un punto cuya coordenada en el eje de abscisas en el tamaño de muestra del estudio. Los valores representados entre los límites de confianza muestran un resultado quirúrgico

adecuado, mientras que un resultado por encima del límite de alarma superior puede ser interpretado como una actividad quirúrgica deficiente.

En Cirugía Cardiovascular se han utilizado multitud de modelos, aunque solo unos pocos han sido aceptados de forma colectiva. Los modelos actualmente más empleados son el de la STS (*Society of Thoracic Surgeons*), el principal en Norteamérica, y EuroSCORE (*European System for Cardiac Operative Risk Evaluation)*, el más extendido en Europa.

STS *Score* fue diseñado como una recogida continua de datos peroperatorios desde su inicio en 1989. Actualmente cuenta con datos de más de 4.500.000 de pacientes intervenidos de cirugía cardiaca procedentes de más de 250 hospitales de Norteamérica. El modelo predictivo es renovado periódicamente mediante reanálisis de la serie completa, lo que supone un excelente método de reajuste, completándose su última actualización en 2009.

EuroSCORE fue desarrollado con datos recogidos en 1995. Ciento treinta y dos centros de ocho países europeos participaron en el proyecto. El modelo fue presentado en 1999, ganando popularidad inmediatamente tras su publicación. Los estudios iniciales de validación externa confirmaron una excelente habilidad predictiva en diferentes ámbitos geográficos, culturales y sociales, subgrupos concretos de patología e incluso para técnicas quirúrgicas que no fueron incluidas en el estudio original. Sin embargo, EuroSCORE nunca fue validado en España mediante un estudio multicéntrico. Solamente un análisis por países de la base de datos original resaltó hallazgos importantes, como una mayor mortalidad observada, atribuida a una casuística más compleja y una mayor proporción de cirugía no coronaria.

A pesar del buen recibimiento inicial, la aplicabilidad de EuroSCORE comenzó a ser cuestionada cuando diversas publicaciones demostraron fallos de validación y sobreestimación del riesgo quirúrgico. Además, EuroSCORE demostró un bajo valor predictivo en el grupo de pacientes de alto riesgo, emergente candidato

para el tratamiento de la valvulopatía aórtica mediante técnicas transcatéter. Los estudios más importantes sobre esta técnica tuvieron que adaptar sus criterios de inclusión debido a esas limitaciones del modelo.

EuroSCORE se había convertido básicamente en un modelo desfasado, anclado en un concepto de cirugía ya pasado. EuroSCORE II fue publicado en 2012, y construido de forma similar sobre las bases de la versión previa con un nuevo conjunto de datos que incluía más de 20.000 pacientes de 154 centros y 43 países. Los datos se recopilaron durante 2010 y las variables fueron reestructuradas mediante adición, eliminación o redefinición.

Un cambio fundamental de EuroSCORE II respecto al original es la definición de mortalidad empleada, que se define como la estrictamente intrahospitalaria. La razón subyacente para este cambio fue la pérdida en el seguimiento a 30 y 90 días, como reflejo de una base de datos deficitaria. El nuevo modelo fue criticado por la existencia de una tendencia a la infraestimación de la mortalidad en cirugía cardiaca comparada con la real. Este hecho despertó dudas por sus implicaciones en los estudios de control de la calidad. La publicación original de EuroSCORE II fue duramente criticada en un editorial, indicando a los lectores que no utilizasen el modelo para estudios científicos o de control de calidad. Los estudios de validación han obtenido resultados contradictorios, con una amalgama de buenos resultados y fallos de calibración. Los artículos publicados en España muestran diferencias importantes entre la mortalidad esperada y observada, incluyendo un estudio multicéntrico.

La relación volumen-resultado ha sido descrita en numerosos estudios, aunque su caracterización es difícil y se ve influida por múltiples variables. En la práctica quirúrgica se ha admitido que los centros con mayor actividad obtienen mejores resultados, expresados como baja mortalidad y complicaciones. EuroSCORE II fue diseñado como una recogida de datos durante un periodo temporal cerrado. Este diseño propició que los centros de mayor volumen aportaran un mayor número de pacientes al modelo, por lo que es esperable que el rendimiento se

vea más afectado por los resultados de estos centros antes que por unidades de menor volumen, que no contribuyeron a su desarrollo en los mismos términos.

La hipótesis de este estudio es que los centros de medio volumen están pobremente representados en el diseño de EuroSCORE II, por lo que es esperable un fallo de validación en este medio.

Métodos

Diseño del estudio

Se diseñó un estudio ambispectivo y longitudal. Los datos fueron obtenidos de la actividad quirúrgica del Hospital General Universitario de Alicante. Aproximadamente la mitad de los pacientes fueron reclutados de forma retrospectiva, mientras que el resto se recogió de forma prospectiva. Este centro tuvo una actividad promedio de entre 300 y 350 intervenciones de cirugía cardiaca mayor en los años previos al inicio del estudio. De acuerdo con los datos nacionales e internacionales disponibles, que informan de una actividad media de entre 250 y 350 intervenciones anuales por centro en España, decidimos clasificar al Servicio del estudio como de medio volumen.

Objetivos del estudio

El tamaño de la muestra y el periodo de inclusión fueron calculados en base a la actividad durante los años previos, con el objetivo de registrar entre 90 y 100 muertes peroperatorias. Este número fue considerado suficiente para alcanzar una potencia estadística adecuada para la prueba de calibración.

Obtención de datos

Fueron incluidos todos los pacientes sometidos a cirugía cardiaca mayor (definida como cualquier intervención quirúrgica realizada para el tratamiento de la patología estructural del corazón y/o grandes vasos torácicos, requiriendo apertura del tórax) durante el periodo de inclusión. Los criterios de exclusión fueron edad menor de 17 años, cirugía cardiaca menor (habitualmente cirugía de

marcapasos sin circulación extracorpórea), cirugía de la pared torácica, implantación de válvulas transcatéter, asistencias circulatorias y trasplantes, ya que dichos procedimientos no fueron incluidos en el original de EuroSCORE II. Las reintervenciones por cualquier motivo dentro del mismo ingreso de la intervención que lo originó también fueron excluidas. Asimismo se descartaron aquellos pacientes intervenidos en 2010 cuyos datos fueron remitidos para el proyecto de desarrollo de EuroSCORE II.

Las variables recogidas fueron todos los datos necesarios para el cálculo de EuroSCORE y EuroSCORE II, así como la mortalidad observada. La mortalidad esperada fue calculada exclusivamente por el autor. La mortalidad fue definida como estrictamente intrahospitalaria, de acuerdo con la definición de EuroSCORE II. Los datos fueros procesados y almacenados en una hoja de cálculo de Microsoft Excel ® (Microsoft Corporation, Redmond, WA, EEUU) diseñada por el autor. Los registros fueron obtenidos de diversas fuentes hospitalarias, y fueron comprobados y depurados hasta obtener cero datos perdidos o incoherente. El conjunto final de datos fue volcado a un archivo de SPSS ® (International Business Machines Corporation, Armonk, NY, EEUU) para su análisis estadístico.

Análisis estadístico

No se realizó ningún análisis por subgrupos. El rendimiento del modelo fue evaluado mediante el cálculo de la calibración y la discriminación. Ésta última fue evaluada mediante el cálculo del área bajo la curva ROC con un intervalo de confianza al 95%. Para el cálculo de la calibración se empleó la prueba de bondad de ajuste de Hosmer-Lemeshow (estadístico de contraste y valor de p). Se construyó un gráfico de embudo para la evaluación de la mortalidad ajustada al riesgo, utilizando EuroSCORE como modelo de referencia para la mortalidad esperada, y siguiendo la metodología descrita por Spiegelhalter. El intervalo de confianza al 99% y la mortalidad de referencia fueron obtenidos del Primer Informe del Proyecto de Calidad de Cirugía Cardiaca del Adulto, siendo el tamaño de muestra el parámetro de precisión. Se realizaron análisis descriptivos de las variables del modelo, expresando las variables continuas como media ±

desviación estándar, y la mortalidad y resto de variables categóricas con medidas de frecuencia (tantos por ciento).

Resultados

Los registros de los años 2010 y 2011 fueron obtenidos retrospectivamente, mientras que los datos de los pacientes intervenidos en 2012 y 2013 se obtuvieron de forma prospectiva. Se incluyeron 970 pacientes en el estudio entre enero de 2010 y agosto de 2013, con una tasa final de cero datos perdidos o incoherentes. Previamente fueron descartados 71 pacientes del año 2010, incluidos en el desarrollo de EuroSCORE II. A pesar de que el proyecto original incluía todos los pacientes intervenidos en los años 2010 y 2013, el reclutamiento de pacientes finalizó en Agosto de 2013 debido al cese de la relación laboral del autor con el centro del estudio.

Los datos referentes a la prevalencia de los factores de riesgo están resumidos en la Tabla 4. La mortalidad observada fue de 82 pacientes (8,5% ± 2,8%). La mortalidad esperada media según EuroSCORE II fue del 3,6% ± 4,1%.

El gráfico de embudo mostró un rendimiento adecuado de la unidad quirúrgica del estudio. La mortalidad de referencia fue del 5,7%, mientras que la mortalidad ajustada al riesgo calculada fue del 7,2%. El punto que representa este dato se encuentra situado dentro de los límites de confianza, aunque con un pequeño y no significativo aumento de la mortalidad que lo sitúa cercano al límite superior de alarma (Figura 5).

La prueba de bondad de ajuste para la calibración obtuvo un resultado de 13,6 (p=0,09). Este resultado refleja un diferencia estadísticamente no significativa, aunque los valores crudos son notablemente diferentes, con una relevancia clínica importante. Las tablas de datos mostraron infraestimación del riesgo quirúrgico en los deciles de alto riesgo, mientras que se obtuvo sobrestimación del riesgo en deciles bajos (Tabla 5).

La curva ROC demostró una capacidad discriminativa adecuada para el modelo. El estadístico *c* fue de 0,72 (intervalo de confianza al 95%: 0,67-0,78) (Figura 5).

Discusión

Este estudio es relevante debido a la necesidad de realizar validaciones externas de los modelos predictivos, así como por el controvertido escenario actual que rodea los resultados de validaciones previas de EuroSCORE II. El Proyecto de Validación en España, el más importante en el país, llevado a cabo por García-Valentín et al., mostró un problema de infraestimación del riesgo quirúrgico en este modelo, aunque menor que la sobrestimación descrita para la versión anterior.

El diseño del proyecto de validación que ahora nos ocupa es adecuado; el tamaño de la muestra fue calculado de forma apropiada, aunque el registro de datos no fue completo debido al descenso de actividad quirúrgica durante 2012 y la pérdida de acceso a los datos del último cuatrimestre de 2013. No hubo datos perdidos, lo cual refleja una gran calidad en el proceso de recogida y depuración. De todo lo anterior se puede afirmar que los datos generados en España son de gran calidad por su grado de compleción y análisis en los últimos estudios realizados.

Los pacientes sometidos a cirugía cardiaca en España han mostrado persistentemente un perfil de alto riesgo en estudios previos. Sin embargo este hecho no ha sido tan marcado en este proyecto. Si comparamos estos datos con los del Proyecto de Validación en España, en nuestra muestra podemos encontrar una menor proporción de pacientes con endocarditis activa, estado preoperatorio crítico, fracción de eyección severamente deprimida o en clase funcional avanzada. Los resultados del proyecto de Validación de EuroSCORE II en España han podido estar influidos por la heterogeneidad de las fuentes, al tratarse de un estudio multicéntrico, al contrario que en este proyecto, donde todos los datos y

cálculos han sido realizados por el autor, eliminando así el error entre observadores.

La mortalidad observada fue del 8,5%, que es discretamente superior a los valores obtenidos en otros estudios realizados en España. El gráfico de embudo mostró que la mortalidad ajustada al riesgo se encontraba entre los límites del intervalo de confianza del gráfico, con una tendencia no significativa al aumento de mortalidad comparado con la de referencia, que podría estar algo infraestimada por el (en la práctica inevitable) sesgo de participación voluntaria. Si obtenemos la mortalidad de referencia y el resto de parámetros del proyecto de Validación en España, el resultado del gráfico de embudo es superponible al anterior. Este pequeño aumento de la mortalidad comparativa concuerda con la hipótesis de este trabajo, por la que los centros de menor volumen obtendrían peores resultados, aunque en este estudio en concreto el periodo de estudio probablemente no es representativo de la actividad global del Servicio, como se comentará más adelante. La referencia empleada para el cálculo de la mortalidad esperada fue la versión original de EuroSCORE, la única validada hasta el momento en España. El Proyecto de Validación en España concluyó que éste probablemente ya no sea el mejor modelo para este propósito, ya que se confirmó el problema de sobrestimación. A pesar de ello esta conclusión se obtuvo de una muestra de datos agrupados procedentes de varios centros con diferente carga de volumen, por lo que no podemos afirmar su aplicabilidad en nuestro caso concreto, el centro de medio volumen.

La discriminación obtuvo un resultado aceptable (0,72), con un intervalo de confianza estrecho y suficientemente alejado del valor de fallo (0,5). Este resultado es superponible a los de otros estudios realizados en España (aunque ligeramente peores), y muy similares al valor que presentó EuroSCORE en la misma muestra.

La prueba de calibración no obtuvo una diferencia estadísticamente significativa, con un valor de p de 0,09, a pesar de que la mortalidad observada fue más del

doble de la esperada. El nivel de significación estadística (error α) habitualmente empleado es 0,05, valor muy cercano al obtenido. A pesar de que los cálculos iniciales eran adecuados, la causa más probable de este hecho es que el tamaño de la muestra del estudio no aportó suficiente potencia estadística para el estudio. El tamaño de la muestra es complejo de calcular en las prueba de calibración, ya que depende en cierta medida de la frecuencia de aparición del evento resultado. Generalmente se acepta que son necesarios 100 eventos (muertes peroperatorias) para lograr una valoración adecuada de la calibración. En este trabajo obtuvimos 82 fallecimientos durante el periodo de reclutamiento, dato cercano al objetivo, pero ligeramente insuficiente. Se puede estimar una potencia estadística del 40% con estos resultados, cuando se recomienda generalmente que sea del 80%, por lo que una muestra ligeramente mayor habría obtenido con probabilidad una diferencia estadísticamente significativa.

Por otro lado, algunos autores opinan que el valor de p no es el único parámetro a considerar en los resultados de un estudio, ya que la magnitud y relevancia del efecto son igualmente importantes. Asimismo el valor de p no puede ser considerado como una valor de todo o nada en la prueba de Hosmer-Lemeshow. Su resultado debe ser interpretado como mejor calibración cuando más cercano sea el valor de p a uno, y peor cuanto más lo sea a cero.

Los resultados de este trabajo son concordantes con otros estudios realizados en España, aunque muchos trabajos realizados en otras partes del mundo han mostrado conclusiones contrarias y conflictivas. Esta situación de un casi fallo de calibración por infraestimación del riesgo puede ser explicada por un pequeño aumento real de la mortalidad quirúrgica y por defectos del modelo de predicción. La población de pacientes que se intervienen en España siempre ha mostrado un perfil de alto riesgo a lo largo del tiempo, probablemente debido a la características especiales del sistema sanitario nacional. No creemos que este proyecto de validación contenga problemas metodológicos, a excepción de la ya mencionada pérdida de potencia estadística. En realidad este estudio tiene la ventaja de contar con la aplicación estricta de las definiciones de las variables y la

eliminación del error entre observadores, ya que el autor ha sido la única persona que ha gestionado los datos. La calidad de los mismos es óptima, con cero registros perdidos o incoherentes, en la misma línea de otros proyectos nacionales.

Este estudio apoya la existencia de una relación volumen-resultado en Cirugía Cardiovascular. Los modelos predictivos como EuroSCORE II, principalmente desarrollados con datos de centros de alta carga quirúrgica, muestran una pequeña infraestimación en centros de medio volumen (o bien esas unidades de medio volumen tienen un ligero aumento de la mortalidad comparadas al modelo de predicción). Probablemente la versión original de EuroSCORE concuerde mejor con los resultados de Servicios más pequeños, como se puede apreciar en el gráfico de embudo. Es posible que la nueva versión no encaje con los resultados de centros de medio volumen al estar peor representados en el modelo. Una posible estrategia para mejorar la relación volumen-resultado en centros medios y pequeños sería la de incrementar la proporción intervenciones/cirujano mediante la reducción de las plantillas, así como la creación de unidades especializadas para los cuidados postoperatorios, de forma que todos los equipos asistenciales puedan adquirir mayor experiencia.

Los resultados de esta validación contribuyen a acrecentar la controversia suscitada por la nueva versión de EuroSCORE, aunque son congruentes con las conclusiones de otras validaciones realizadas en el mismo país. El equipo de desarrollo de EuroSCORE II sugirió realizar validaciones del modelo en cada centro que realice cirugía cardiaca de forma individual. Con ello se evaluaría el rendimiento de cada unidad con respecto al modelo, creando así un cociente de mortalidad ajustada al riesgo, que debería ser multiplicado al riesgo predicho por EuroSCORE II para obtener el riesgo real del paciente en ese centro concreto. Por último, es importante destacar que los resultados y conclusiones de este estudio deben ser aplicados a la actividad global de un Servicio, ya que no ha sido diseñado para subgrupos concretos (cirugía valvular, coronaria, etc.).

La principal limitación de este estudio es la insuficiente potencia estadística que no pudo mostrar una diferencia estadísticamente significativa en la prueba de calibración, aun cuando existe una evidente e importante diferencia entre la mortalidad esperada y la observada. La causa de esto fue la pérdida de pacientes por una doble razón, el descenso de la actividad quirúrgica durante 2012 y la imposibilidad de recolectar los datos del último cuatrimestre de 2013 debido al cese de la actividad del autor en dicho centro. Además se descartaron 71 pacientes para realizar una validación externa de forma rigurosa, ya que sus datos fueron remitidos para el desarrollo de EuroSCORE II. Si añadimos todos estos pacientes, la muestra aumentaría hasta aproximadamente 1.175 pacientes, con una alta probabilidad de alcanzar el objetivo de 100 fallecimientos peroperatorios, que probablemente habrían otorgado una adecuada potencia estadística. La recogida retrospectiva de los datos de los años 2010 y 2011 es otra de las limitaciones, ya que ésta genera pérdida de datos e infraestimación de la mortalidad esperada.

La variabilidad de la mortalidad en el Servicio durante el periodo de inclusión es un aspecto que pudo afectar a los resultados, especialmente el del gráfico de embudo. Coincidiendo con el descenso de la actividad quirúrgica durante 2012 se apreció un aumento relevante de la mortalidad quirúrgica, sin apreciarse un incremento importante del riesgo operatorio que pudiese explicarlo. Si igualamos los resultados de 2012 con los del resto del periodo de inclusión, la mortalidad observada habría sido sustancialmente menor, ajustándose mejor a la mortalidad de referencia nacional mencionada previamente.

Las expectativas futuras de esta línea de investigación deberían pasar por la realización de un estudio multicéntrico en unidades de pequeño y medio volumen, de forma que se pudiese obtener una potencia estadística y homogeneidad adecuadas para la obtención de resultados concluyentes, incluso utilizando las dos versiones del modelo para averiguar cuál de las dos se ajusta mejor a este tipo de centros. Actualmente el autor está proponiendo un nuevo estudio para

analizar las diferencias en las prevalencias de los entre la población española y la base de datos de desarrollo de EuroSCORE II.

Conclusiones

La discriminación de EuroSCORE II es adecuada en un centro de medio volumen. La prueba de bondad de ajuste muestra una diferencia relevante, aunque no estadísticamente significativa por una baja potencia estadística.

La mortalidad quirúrgica en el Servicio de Cirugía Cardiaca del Hospital General Universitario de Alicante no difiere de forma significativa con la media española.

La distribución y frecuencias de los factores de riesgo son similares a las de otros estudios españoles, aunque con un perfil de menor riesgo global.

8.2. Inglés

Introduction

Evaluation of risk is a fundamental part of the decision-making process for every patient in whom cardiac surgery is otherwise indicated. Predictive models are used worldwide to aid this process, as they integrate patient and procedural variables, and aim to assimilate this data to estimate the specific risk of mortality. They can be constructed with retrospective and prospectively collected registries, and use administrative or clinical datasets. Probably prospective studies carried on clinical data are the best methodology for this purpose.

The most important issue for predictive model creation is an adequate selection of the variables. Predictive variables (predictors) must have a high prevalence and statistic association with the outcome variable, and their definition must be strict. The results must be also relevant and well defined, which are called *hard* results. The most extended outcome in surgical predictive models is mortality. Data loss is

frequent in predictive variables, although they can sometimes be completed with imputation techniques. However, loss of outcome data is a very serious problem, which can hardly be repaired.

Predictive models are constructed employing regression methods. The most common is the binary logistic regression, used when the outcome is a discontinuous and dichotomic variable. The are different types of predictors (continuous or not, dichotomic, polytomic) that have to be managed with different mathematical approaches, although the most frequent practice is to transform as much as possible predictors is dichotomic (yes or not) answers.

There are specific problems related to predictive models design, which require a watchful observation. Overfitting is generated when the model has a good performance in the developmental dataset, but fails to do it in the general population. The difference between both validation results is called optimism, and is usually observed when analysis of multiple predictors is required. Model generation can be also affected by other biases, as selection, observation or publication. The last concept handled in these studies is called parsimony, and reflects the idea of constructing the simplest working model.

Predictive models need also to be tested externally in validation studies. The most common methodology, among diverse statistic tests, is to evaluate the calibration (mean expected mortality matches with observed) and discrimination (patients in higher risk are those who actually die) as different aspects. The most popular tests are the (currently discussed) Hosmer-Lemeshow goodness-of-fit test for calibration and the area under the ROC (Receiver Operating Characteristics) curve for discrimination. A key issue in validation studies is to calculate the optimal sample size, which is especially difficult for the calibration test. One hundred result events are usually accepted as adequate for this purpose. Calibration failures can be explained by several reasons; model deficiencies, actual poor outcomes or validation mistakes. A careful analysis of the causes is mandatory in these situations.

Funnel plots with 99% confidence intervals are statistical and graphic tools used for quality studies. They show a mortality reference as a horizontal line crossing the Y-axis and two alarm limits that narrow around it as sample size becomes larger (with the characteristic funnel shape). Risk-adjusted mortality is plotted as a dot over the sample size in the X-axis. A value between the limits show adequate surgical results of the centre, while mortalities above the upper bounds show a risk-adjusted mortality over the expected reference that can be interpreted as a bad surgical performance.

Many predictive models have been introduced and used in cardiac surgery, although only a few have ever been widely accepted or adapted. The most widely used models currently are the STS (Society of Thoracic Surgeons) Score, the main score applied across North America, and EuroSCORE (European System for Cardiac Operative Risk Evaluation), which is the most commonly used model throughout Europe.

The STS Score was designed to utilise continuously collected perioperative data, which began in 1989. It currently includes data from over 4.500.000 patients enrolled in registries of more than 250 hospitals in North America, who underwent major cardiac operations. The predictive model based on this dataset is periodically updated following re-analysis of the whole series. This is an excellent method for readjustment, and the latest update was performed in 2009.

EuroSCORE was developed with data originally collected in 1995. One hundred and thirty-two centres from eight European countries participated in the project. The model developed was published in 1999 and became popular shortly after its publication. External validation studies confirmed a good predictive ability in different geographical, social and cultural populations, specific subgroups of disease or even for operative techniques that had not been included in the original dataset. However EuroSCORE has never been specifically validated in a multicentre study in Spain. Only a country-based analysis highlighted important

findings, such as a high actual observed mortality, which was attributed to a more complex case-mix, and a higher proportion of non-coronary surgery in the population.

Although initially well received, questions about the applicability of EuroSCORE began to be raised as publications demonstrated validation failures and overestimation of the mortality risk. In addition, EuroSCORE was found to have poor predictive value in the high-risk patient group, who emerged as potential candidates for aortic valve surgery or transcatheter or aortic valve implantation. The most significant studies on transcatheter aortic valve implantation had to adapt their inclusion criteria due to these limitations in EuroSCORE.

EuroSCORE had essentially become an outdated model, relevant to cardiac surgery of a bygone period. EuroSCORE II was published in 2012, and based on the previous version, it was constructed in the same way, with a new dataset including over 20.000 patients from 154 centres and 43 countries worldwide. Data were collected during 2010 and predictive variables were restructured by adding, eliminating or redefining them.

A fundamental change of EuroSCORE II compared with the original model is the definition of mortality, strictly predicting in-hospital mortality rate. The underlying reason for this change was the loss of follow-up data at 30 and 90 days, reflecting a low quality dataset. The revised model was criticised for a trend to underestimation of mortality from cardiac surgery compared with the actual mortality figures. This raised concerns also as it held implications for quality control studies. The original publication of EuroSCORE II was hardly criticised in an editorial letter, suggesting to readers to not to use it for scientific purpose or quality control. Validation studies have shown contradictory results with a mix of good performance and miscalibration. Papers published in Spain have shown significant differences between expected and observed mortalities, including a prospective and multicentre study.

Volume-outcome relationship has been described in diverse studies, although its characterisation is difficult and influenced by multiple variables. In surgery, it is globally admitted that centres with higher surgical activity report better outcomes, expressed as low postoperative mortality or complications. EuroSCORE II was projected as a data recruitment during a limited time period. This design allowed surgical units with a larger volume to provide a higher number of patients to this model. That way, it is expected that model performance is mostly influenced by the results of these centres, but other units with medium or low volume did not contribute in the same way.

The hypothesis of this study is that medium volume centres are poorly represented in EuroSCORE II design, so a validation failure can be expected in this scenario.

Methods

Study design

An ambispective, longitudinal study was designed. Data were recruited from the surgical activity of the University General Hospital of Alicante. Proximately a half of patients were recorded retrospectively, and the second half prospectively. This centre had an average activity between 300 and 350 major cardiac interventions in previous years. According to national and international data, average Spanish activity was set in a number between 250 and 350 interventions per year, so we decided to classify the study service as a medium volume centre.

Study objectives

The primary objective of the study was to evaluate the external performance of EuroSCORE II in the Department of Cardiovascular Surgery of the University General Hospital of Alicante.

As secondary endpoints, quality assessment was performed using an original EuroSCORE based funnel plot, and a descriptive analysis of risk factors was also performed.

Sample size

Sample size and inclusion period were estimated according to the surgical activity of prior years and the mortality rates from national and own data, aiming to reach a number of 90-100 postoperative deaths. This number should be sufficient to reach an adequate statistical power for the calibration test.

Data collection

All patients undergoing major cardiac surgery (defined as surgical intervention for treatment of a structural heart disease or thoracic large vessels impairment, requiring opening of the thorax) during the inclusion period were included. Exclusion criteria were patients under 17 years, minor cardiac surgery (usually pacemaker surgery not requiring cardiopulmonary bypass), chest wall surgery, transcatheter valve implantation, mechanical circulatory support implantation and transplants, as these procedures were not included in the original EuroSCORE II dataset. Reinterventions for any cause in the same admission as the primary operation were also excluded. Patients operated in 2010 whose data were also sent for the development of EuroSCORE II were discarded.

Variables collected were all the necessary data for EuroSCORE and EuroSCORE II calculation and observed mortality as result variable. Expected mortality was exclusively calculated by the author. Mortality was defined as in-hospital, the same definition employed in EuroSCORE II model. A Microsoft Excel ® (Microsoft Corporation, Redmond, WA, USA) tool was designed by the author for data calculation and storage. Data were obtained from diverse hospital records, and were checked and depurated with the aim of realising a zero incidence of missing or incoherent data. Dataset was loaded in a single SPSS ® (International Business Machines Corporation, Armonk, NY, USA) file for statistical analysis.

Statistical analysis

Subgroup analysis was not performed. Performance of the model was assessed via the measurement of calibration and discrimination. Discrimination was

evaluated with the area under the ROC curve with a 95% confidence interval. Calibration was evaluated with the Hosmer-Lemeshow goodness-of-fit test (statistic and p values). A funnel plot was constructed for risk-adjusted mortality assessment, using EuroSCORE as reference for expected data, and following the Spiegelhalter methodology. Confidence interval 99% and reference mortality (5,7%) was extracted from the First Report of the Quality Project of Adult Cardiac Surgery in Spain, with the sample size as precision parameter. Descriptive analyses were performed for every predictive variable of the model, expressing continuous variables with the mean ± standard deviation, and mortality and other categorical variables with frequency measures (percentages).

Results

Records from 2010 and 2011 were obtained retrospectively, while the data from patients operated in 2012 and 2013 were recruited prospectively. From January 2010 to August 2013, 970 patients were included in the studio, with a zero rate of lost or incoherent data. Seventy-one patients from 2010, included in the development of EuroSCORE II have been previously discarded. The initial project design included all patients from 2010 to 2013. Recruitment was stopped in August 2013 due to the cessation of the author's activity in the study centre.

Data regarding prevalence of risk factors and predictive variables are collected in Table 4. The observed mortality was 82 patients (8,5% ± 2,8%). Mean predicted mortality by EuroSCORE II was 3,6% ± 4,1%.

The funnel plot demonstrated an adequate surgical performance for this unit. Reference published mortality was 5,7%, while risk-adjusted mortality was 7,2%. The dot plotting this result stayed between the boundaries, although with a slight trend to a non significant increase in observed mortality that placed it close to the upper limit of the confidence interval (Figure 5).

Calibration goodness-of-fit test had a result for the test of 13,6 (p=0,09). This reflects a non statistically significant difference, although crude values are notably different, with a remarkable clinical relevance. Crosstabs showed underestimation in high-risk deciles and overestimation in low-risk deciles for the model (Table 5).

ROC curve showed an adequate discriminative ability for the model. C-statistic was 0,72 (95% confidence interval 0,67-0,78) (Figure 5).

Discussion

This study is considered as relevant by the necessity of testing external performance of predictive models, and the current scenario of controversy surrounding the results of previous validation studies of EuroSCORE II. The most important Validation Project in Spain was conducted and published by García-Valentín et al., showing a problem of underestimation of operative risk of this model, although smaller than the overprediction detected in the previous version.

The design of this validation project is adequate; sample size was appropriately calculated, although recruitment was not complete due to the drop in the surgical activity during 2012 and the loss of access to the data from the last forth months of 2013. Missing data was zero thus reflecting high quality, scrutiny and collection. From all of the above it can be said that Spanish data are of good quality as from the completeness and analysis of the records in the last performed studies.

Cardiac surgical patients in Spain have shown a high-risk profile in previous studies. In this project, however, risk profile was not so high. When compared with the data from the Validation Project in Spain, we can find a lower rate of patients presenting to surgery on active endocarditis, critical status, very poor ejection fraction or advanced functional class in this sample. Results in the Validation Project in Spain project could also be influenced by the heterogeneity of the observers in a multicentre study, as opposed to this project, where all data and

calculations were performed by the author, thus eliminating the inter observer error.

Observed mortality was 8,5%, that is slightly higher than the values obtained in other Spanish studies. The funnel plot showed the risk-adjusted mortality laying between the confidence limits of the plot, with a non significant trend to the increase of mortality when compared with the reference, although it can be somewhat underestimated due to the (actually unavoidable) voluntary participation bias. If we use the reference mortality and rest of parameters from the Validation Project in Spain, the funnel plot shows the same result. This small rise of mortality matches with the hypothesis of this study, with smaller centres obtaining worse results, although the study period could not be representative of the unit overall activity, as will be further commented. The reference used to calculate expected mortality was the original version of EuroSCORE, as it was the only validated model in Spain. Validation Project in Spain concluded that probably this is no longer a good reference model, so an overestimation problem was detected. This conclusion was obtained from pooled data recruited from diverse centres with different volume load, so we do not know its applicability in our specific scenario of a medium volume unit.

Discrimination result was acceptable (0,72), with a narrow confidence interval far enough from the failure (0,5). This result is consistent with other studies performed in Spain (although somewhat worse), and very similar to the discrimination value shown by EuroSCORE in the same sample.

The calibration test did not show a statistically significant difference, with a p-value of 0,09, although observed mortality is more than twice the expected. The level of significance (α error) usually accepted is 0,05, which is quite close to the obtained p-value. Perhaps the sample size in this study did not provide enough statistical power for the objective, although initial calculations did so. Sample size is difficult to calculate for the calibration test, as it also depends on the frequency of the outcome event. It is commonly accepted that 100 events (perioperative

deaths) are needed to make a good calibration assessment. During the recruitment period 82 events were recorded, which is near the objective, but slightly insufficient. A statistical power of 40% can be estimated with the results obtained in this study, when the recommended value is 80%, so a slightly larger sample should have probably shown a calibration failure.

On the other hand, some authors indicate p-value is not the only parameter to be evaluated in the results, but magnitude and relevance of the effect are equally important. P-value cannot be considered an all-or-nothing concept in the Hosmer-Lemeshow test. Its results can be interpreted as better calibration as closer the p-value is to one, and worse as it goes near zero.

Results in this project are consistent with other works performed in Spain, although many works across the world have shown inconsistent and even conflictive conclusions. The situation of almost a calibration failure by underestimation of risk can be explained by a small but true increase in mortality and deficiencies in the predictive model. Spanish cardiac population has always shown a high-risk profile along time, probably explained by the special characteristics of the National Healthcare System. We do not consider this validation project contains methodological problems, except for the above mentioned loss of statistical power. Actually, this study has the advantages of a strict application of predictors definitions and the elimination of inter observer error, as the author is the only person involved in data management. Data quality is also optimal, with zero lost or incoherent records, on the line of previous national projects.

This study supports the existence of a volume-outcome relationship in cardiac surgical practice. Predictive models as EuroSCORE II, mostly developed with high load centres, show a little underestimation in medium volume centres (or these units show an increase of mortality when compared to the predictive model). The original version of EuroSCORE probably better matches with the results of smaller services, as the funnel plot shows. The new version probably does not fit the

results, as medium volume centres are not well represented in it. A possible strategy to improve the volume-outcome relationship is to increase the interventions/surgeon rate by decreasing the number of staff surgeons, and to create specialized units for postoperative care, so all attending teams can acquire more experience.

The results of this validation contribute to the controversy scenario regarding the new version of EuroSCORE, although they are consistent with the conclusions of other validations in the same country. One of the suggestions of the EuroSCORE II developmental team was to perform individual validations of the model in each centre performing cardiac surgery. A risk-adjusted mortality ratio should be calculated that way, which should be multiplied to the result of EuroSCORE II in order to calculate the actual risk of mortality in that specific centre. Finally, the results and conclusions of this study must be applied only to the overall activity of a surgical team, as it has not been designed for specific subgroups (coronary, valve surgery, etc.).

The main limitation of this study is the insufficient statistical power that could not obtain a statistically significant difference in the calibration test, when there is an evident and relevant difference between expected and observed mortality. The reason for that is a loss of patients due to the drop of the surgical activity during 2012 and the impossibility to recruit the data from the last four months of 2013 due to the author's cessation of activity in the centre. 71 patients were also discarded to maintain the strictness of an external validation, as they were submitted for the EuroSCORE II development. Adding the lost patients, a sample of 1.175 patients is estimated, with a higher likelihood to reach the number of 100 perioperative deaths, which should probably provide an adequate statistical power. Restrospective collection of 2010 and 2011 data is another limitation, as it is known to generate loss of data and underestimation of surgical risk.

Mortality variability during he period is an important issue that could affect the results, especially of the funnel plot. Concurring with the drop in the surgical

activity, observed mortality significantly increased during 2012 only, with no important increase in perioperative risk that could explain it. Matching the results of 2012 with that from the rest of the inclusion period, observed mortality should have adjusted even better to the reference nation-wide mortality.

The future aspects of these investigations should go through the performance of a multicentre study in medium volume centres, to get enough statistical power and homogeneity of the sample to obtain a concluding result, even using both versions of the model to find out which of them better fits these centres results. Nowadays, the author is launching a new study to analyse the differences in risk factor prevalences between Spanish population and the developmental dataset of EuroSCORE II.

Conclusions

Discrimination of EuroSCORE II is adequate in a medium volumen centre. The goodness-of-fit test shows a relevant difference, although not statistically significant due to a low statistic power.

Surgical mortality of the Cardiovascular Surgey Department of the University General Hospital of Alicante is not significanlty different of the Spanish average.

Distribution and frequency of the risk factors are similar to other Spanish studies, although with a lower overall risk profile.

9. REFERENCIAS BIBLIOGRÁFICAS

1. Burgueño MJ, Garcia-Bastos JL, Gonzalez-Buitrago JM. ROC curves in the evaluation of diagnostic tests. Med Clin (Barc). 1995; 104:661-70.

2. Steyerberg E. Clinical Prediction models. 1 ed. New York: Springer Science+Business Media, LLC; 2010.

3. Cortina JM. Scores de gravedad y complejidad en cirugia cardiaca. Usos y limitaciones. Rev Esp Cardiol. 2005; 58:473-6.

4. Suman A, Barnes DS, Zein NN, Levinthal GN, Connor JT, Carey WD. Predicting outcome after cardiac surgery in patients with cirrhosis: a comparison of Child-Pugh and MELD scores. Clin Gastroenterol Hepatol. 2004; 2:719-23.

5. Arif R, Seppelt P, Schwill S, Kojic D, Ghodsizad A, Ruhparwar A, et al. Predictive risk factors for patients with cirrhosis undergoing heart surgery. Ann Thorac Surg. 2012; 94:1947-52.

6. Jacob KA, Hjortnaes J, Kranenburg G, de Heer F, Kluin J. Mortality after cardiac surgery in patients with liver cirrhosis classified by the Child-Pugh score. Interact Cardiovasc Thorac Surg. 2015; 20:520-30.

7. Modi A, Vohra HA, Barlow CW. Do patients with liver cirrhosis undergoing cardiac surgery have acceptable outcomes? Interact Cardiovasc Thorac Surg. 2010; 11:630-4.

8. 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.

9. Roques F, Nashef SA, Michel P, Gauducheau E, de Vincentiis C, Baudet E, et al. Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. Eur J Cardiothorac Surg. 1999; 15:816-22.

10. Nashef SA, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. EuroSCORE II. Eur J Cardiothorac Surg. 2012; 41:734-44.

11. Sergeant P, Meuris B, Pettinari M. EuroSCORE II, illum qui est gravitates magni observe. Eur J Cardiothorac Surg. 2012; 41:729-31.

12. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol. 2003; 56(5):441-7.

13. Bürger G. Las Aventuras del Barón de Münchhausen. Barcelona: Edhasa; 2003.

14. Campbell JP, Maxey VA, Watson WA. Hawthorne effect: implications for prehospital research. Ann Emerg Med. 1995; 26:590-4.

15. Ulmer FC. The Hawthorne effect. Educ Dir Dent Aux. 1976; 1:28.

16. Lehmann C, Nowak A. The Hawthorne effect: can it be measured and utilized? Br J Anaesth. 2013; 110:658-9.

17. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet. 1991; 337:867-72.

18. Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA. 1990; 263:1385-9.

19. Dickersin K, Min YI. Publication bias: the problem that won't go away. Ann N Y Acad Sci. 1993; 703:135-46.

20. Dickersin K, Min YI. NIH clinical trials and publication bias. Online J Curr Clin Trials. 1993; Doc No 50.

21. 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.

22. Lemeshow S, Hosmer DW, Jr. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol. 1982; 115:92-106.

23. Garcia-Valentin A, Mestres CA. Reply to Collins and Le Manach. Eur J Cardiothorac Surg. 2016; 49:358.

24. Collins GS, Le Manach Y. Knowingly repeating an incorrect and inefficient analysis is flawed logic. Eur J Cardiothorac Surg. 2016; 49:357-8.

25. Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. Stat Med. 2013; 32:67-80.

26. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem. 1993; 39:561-77.

27. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29-36.

28. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983; 148:839-43.

29. Pinna-Pintor P, Bobbio M, Colangelo S, Veglia F, Giammaria M, Cuni D, et al. Inaccuracy of four coronary surgery risk-adjusted models to predict mortality in individual patients. Eur J Cardiothorac Surg. 2002; 21:199-204.

30. Borracci RA, Rubio M, Celano L, Ingino CA, Allende NG, Ahuad Guerrero RA. Prospective validation of EuroSCORE II in patients undergoing cardiac surgery in Argentinean centres. Interact Cardiovasc Thorac Surg. 2014; 18:539-43.

31. Spiegelhalter DJ. Funnel plots for comparing institutional performance. Stat Med. 2005; 24:1185-202.

32. Spiegelhalter D. Funnel plots for institutional comparison. Qual Saf Health Care. 2002; 11:390-1.

33. Spiegelhalter DJ. Mortality and volume of cases in paediatric cardiac surgery: retrospective study based on routinely collected data. BMJ. 2002; 324:261-3.

34. Stark J, Gallivan S, Lovegrove J, Hamilton JR, Monro JL, Pollock JC, et al. Mortality rates after surgery for congenital heart defects in children and surgeons' performance. Lancet. 2000; 355:1004-7.

35. Stark JF, Gallivan S, Davis K, Hamilton JR, Monro JL, Pollock JC, et al. Assessment of mortality rates for congenital heart defects and surgeons' performance. Ann Thorac Surg. 2001; 72:169-74.

36. Tekkis PP, McCulloch P, Steger AC, Benjamin IS, Poloniecki JD. Mortality control charts for comparing performance of surgical units: validation study using hospital mortality data. BMJ. 2003; 326:786-8.

37. Normand SL, Shahian D. Statistical and clinical aspects of hospital outcomes profiling. Statist Sci. 2007; 22:206-26.

38. 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.

39. 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:I3-12.

40. The Society of Thoracic Surgeons. The Society os Thoracic Surgeons National Database: http://www.sts.org/national-database; 2014.

41. Cortina J. ¿Es EuroSCORE II el nuevo patrón como modelo de riesgo en cirugía cardiaca? Uso, aplicación clínica, evaluación y consecuencias. Cir Cardiov. 2013; 20:55-8.

42. Roques F, Michel P, Goldstone AR, Nashef SA. The logistic EuroSCORE. Eur Heart J. 2003; 24:881-2.

43. Roques F, Nashef SA, Michel P, Pinna Pintor P, David M, Baudet E, et al. Does EuroSCORE work in individual European countries? Eur J Cardiothorac Surg. 2000; 18:27-30.

44. Nashef SA, Roques F, Michel P, Cortina J, Faichney A, Gams E, et al. Coronary surgery in Europe: comparison of the national subsets of the European system for cardiac operative risk evaluation database. Eur J Cardiothorac Surg. 2000; 17:396-9.

45. Alvarez M, Colmenero M, Martin P, Prades I, Moreno E, Gonzalez-Molina M, et al. Does the EuroSCORE identify patients at minimum risk of mortality from heart surgery? Rev Esp Cardiol. 2003; 56:682-6.

46. Pitkanen O, Niskanen M, Rehnberg S, Hippelainen M, Hynynen M. Intra-institutional prediction of outcome after cardiac surgery: comparison between a locally derived model and the EuroSCORE. Eur J Cardiothorac Surg. 2000; 18:703-10.

47. Andrade IN, Moraes Neto FR, Oliveira JP, Silva IT, Andrade TG, Moraes CR. Assesment of the EuroSCORE as a predictor for mortality in valve cardiac surgery at the Heart Institute of Pernambuco. Rev Bras Cir Cardiovasc. 2010; 25:11-8.

48. Lafuente S, Trilla A, Bruni L, Gonzalez R, Bertran MJ, Pomar JL, et al. Validation of the EuroSCORE probabilistic model in patients undergoing coronary bypass grafting. Rev Esp Cardiol. 2008; 61:589-94.

49. Mestres CA, Castro MA, Bernabeu E, Josa M, Cartana R, Pomar JL, et al. Preoperative risk stratification in infective endocarditis. Does the EuroSCORE model work? Preliminary results. Eur J Cardiothorac Surg. 2007; 32:281-5.

50. Chen CC, Wang CC, Hsieh SR, Tsai HW, Wei HJ, Chang Y. Application of European system for cardiac operative risk evaluation (EuroSCORE) in coronary artery bypass surgery for Taiwanese. Interact Cardiovasc Thorac Surg. 2004; 3:562-5.

51. Parolari A, Pesce LL, Trezzi M, Loardi C, Kassem S, Brambillasca C, et al. Performance of EuroSCORE in CABG and off-pump coronary artery bypass grafting: single institution experience and meta-analysis. Eur Heart J. 2009; 30:297-304.

52. Youn YN, Kwak YL, Yoo KJ. Can the EuroSCORE predict the early and mid-term mortality after off-pump coronary artery bypass grafting? Ann Thorac Surg. 2007; 83:2111-7.

53. Al-Ruzzeh S, Asimakopoulos G, Ambler G, Omar R, Hasan R, Fabri B, et al. Validation of four different risk stratification systems in patients undergoing off-pump coronary artery bypass surgery: a UK multicentre analysis of 2223 patients. Heart. 2003; 89:432-5.

54. Riha M, Danzmayr M, Nagele G, Mueller L, Hoefer D, Ott H, et al. Off pump coronary artery bypass grafting in EuroSCORE high and low risk patients. Eur J Cardiothorac Surg. 2002; 21:193-8.

55. Nashef SA, Roques F, Hammill BG, Peterson ED, Michel P, Grover FL, et al. Validation of European System for Cardiac Operative Risk Evaluation (EuroSCORE) in North American cardiac surgery. Eur J Cardiothorac Surg. 2002; 22:101-5.

56. Sergeant P, de Worm E, Meyns B. Single centre, single domain validation of the EuroSCORE on a consecutive sample of primary and repeat CABG. Eur J Cardiothorac Surg. 2001; 20:1176-82.

57. Michel P, Roques F, Nashef SA, Euro SPG. Logistic or additive EuroSCORE for high-risk patients? Eur J Cardiothorac Surg. 2003; 23:684-7.

58. Hickey GL, Grant SW, Murphy GJ, Bhabra M, Pagano D, McAllister K, et al. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. Eur J Cardiothorac Surg. 2013; 43:1146-52.

59. Yap CH, Reid C, Yii M, Rowland MA, Mohajeri M, Skillington PD, et al. Validation of the EuroSCORE model in Australia. Eur J Cardiothorac Surg. 2006; 29:441-6.

60. Wang C, Yao F, Han L, Zhu J, Xu ZY. Validation of the European system for cardiac operative risk evaluation (EuroSCORE) in Chinese heart valve surgery patients. J Heart Valve Dis. 2010; 19:21-7.

61. Leon MB, Smith CR, Mack M, Miller DC, Moses JW, Svensson LG, et al. Transcatheter aortic-valve implantation for aortic stenosis in patients who cannot undergo surgery. N Engl J Med. 2010; 363:1597-607.

62. Smith CR, Leon MB, Mack MJ, Miller DC, Moses JW, Svensson LG, et al. Transcatheter versus surgical aortic-valve replacement in high-risk patients. N Engl J Med. 2011; 364:2187-98.

63. Parolari A, Pesce LL, Trezzi M, Cavallotti L, Kassem S, Loardi C, et al. EuroSCORE performance in valve surgery: a meta-analysis. Ann Thorac Surg. 2010; 89:787-93.

64. Kalavrouziotis D, Li D, Buth KJ, Legare JF. The European System for Cardiac Operative Risk Evaluation (EuroSCORE) is not appropriate for withholding surgery in high-risk patients with aortic stenosis: a retrospective cohort study. J Cardiothorac Surg. 2009; 4:32.

65. Qadir I, Salick MM, Perveen S, Sharif H. Mortality from isolated coronary bypass surgery: a comparison of the Society of Thoracic Surgeons and the EuroSCORE risk prediction algorithms. Interact Cardiovasc Thorac Surg. 2012; 14:258-62.

66. Tran DT, Dupuis JY, Mesana T, Ruel M, Nathan HJ. Comparison of the EuroSCORE and Cardiac Anesthesia Risk Evaluation (CARE) score for risk-adjusted mortality analysis in cardiac surgery. Eur J Cardiothorac Surg. 2012; 41:307-13.

67. Zheng Z, Li Y, Zhang S, Hu S, Chinese CRS. The Chinese coronary artery bypass grafting registry study: how well does the EuroSCORE predict operative risk for Chinese population? Eur J Cardiothorac Surg. 2009; 35:54-8.

68. Ranucci M, Castelvecchio S, Menicanti LA, Scolletta S, Biagioli B, Giomarelli P. An adjusted EuroSCORE model for high-risk cardiac patients. Eur J Cardiothorac Surg. 2009; 36:791-7.

69. Silva J, Ridao-Cano N, Segura A, Maroto LC, Cobiella J, Carnero M, et al. Can estimated glomerular filtration rate improve the EuroSCORE? Interact Cardiovasc Thorac Surg. 2008; 7:1054-7.

70. Nozohoor S, Sjogren J, Ivert T, Hoglund P, Nilsson J. Validation of a modified EuroSCORE risk stratification model for cardiac surgery: the Swedish experience. Eur J Cardiothorac Surg. 2011; 40:185-91.

71. Lebreton G, Merle S, Inamo J, Hennequin JL, Sanchez B, Rilos Z, et al. Limitations in the inter-observer reliability of EuroSCORE: what should change in EuroSCORE II? Eur J Cardiothorac Surg. 2011; 40:1304-8.

72. Nashef SA. Applying and evaluating risk models. Eur J Cardiothorac Surg. 2012; 41:314-5.

73. Lin CH, Hsu RB. Cardiac surgery in patients with liver cirrhosis: risk factors for predicting mortality. World J Gastroenterol. 2014; 20:12608-14.

74. Shuhaiber JH, Goldsmith K, Nashef SA. The influence of seasonal variation on cardiac surgery: a time-related clinical outcome predictor. J Thorac Cardiovasc Surg. 2008; 136:894-9.

75. Poullis M, Fabri B, Pullan M, Chalmers J. Sampling time error in EuroSCORE II. Interact Cardiovasc Thorac Surg. 2012; 14:640-1.

76. Hickey GL, Bridgewater B. How well calibrated is EuroSCORE II? Eur J Cardiothorac Surg. 2013; 43:208.

77. Sharples LD, Nashef SA, Euro SPG. Reply to Hickey and Bridgewater. Eur J Cardiothorac Surg. 2013; 43:208-9.

78. Nezic D, Borzanovic M, Spasic T, Vukovic P. Calibration of the EuroSCORE II risk stratification model: is the Hosmer-Lemeshow test acceptable any more? Eur J Cardiothorac Surg. 2013;4 3:206.

79. Sharples LD, Nashef SA, Euro SPG. Reply to Nezic et al. Eur J Cardiothorac Surg. 2013; 43:207.

80. Noyez L, Kievit PC, van Swieten HA, de Boer MJ. Cardiac operative risk evaluation: The EuroSCORE II, does it make a real difference? Neth Heart J. 2012; 20:494-8.

81. Biancari F, Vasques F, Mikkola R, Martin M, Lahtinen J, Heikkinen J. Validation of EuroSCORE II in patients undergoing coronary artery bypass surgery. Ann Thorac Surg. 2012; 93:1930-5.

82. Di Dedda U, Pelissero G, Agnelli B, De Vincentiis C, Castelvecchio S, Ranucci M. Accuracy, calibration and clinical performance of the new EuroSCORE II risk stratification system. Eur J Cardiothorac Surg. 2013; 43:27-32.

83. Chalmers J, Pullan M, Fabri B, McShane J, Shaw M, Mediratta N, et al. Validation of EuroSCORE II in a modern cohort of patients undergoing cardiac surgery. Eur J Cardiothorac Surg. 2013; 43:688-94.

84. Zhang GX, Wang C, Wang L, Lu FL, Li BL, Han L, et al. Validation of EuroSCORE II in Chinese patients undergoing heart valve surgery. Heart Lung Circ. 2013; 22:606-11.

85. Wang L, Han QQ, Qiao F, Wang C, Zhang XW, Han L, et al. Performance of EuroSCORE II in patients who have undergone heart valve surgery: a multicentre study in a Chinese population. Eur J Cardiothorac Surg. 2014; 45:359-64.

86. Kunt AG, Kurtcephe M, Hidiroglu M, Cetin L, Kucuker A, Bakuy V, et al. Comparison of original EuroSCORE, EuroSCORE II and STS risk models in a Turkish cardiac surgical cohort. Interact Cardiovasc Thorac Surg. 2013; 16:625-9.

87. Qadir I, Alamzaib SM, Ahmad M, Perveen S, Sharif H. EuroSCORE vs. EuroSCORE II vs. Society of Thoracic Surgeons risk algorithm. Asian Cardiovasc Thorac Ann. 2014; 22:165-71.

88. Paparella D, Guida P, Di Eusanio G, Caparrotti S, Gregorini R, Cassese M, et al. Risk stratification for in-hospital mortality after cardiac surgery: external validation of EuroSCORE II in a prospective regional registry. Eur J Cardiothorac Surg. 2014; 46:840-8.

89. Osnabrugge RL, Speir AM, Head SJ, Fonner CE, Fonner E, Kappetein AP, et al. Performance of EuroSCORE II in a large US database: implications for transcatheter aortic valve implantation. Eur J Cardiothorac Surg. 2014; 46:400-8.

90. Barili F, Pacini D, Capo A, Ardemagni E, Pellicciari G, Zanobini M, et al. Reliability of new scores in predicting perioperative mortality after isolated aortic valve surgery: a comparison with the society of thoracic surgeons score and logistic EuroSCORE. Ann Thorac Surg. 2013; 95:1539-44.

91. Barili F, Pacini D, Capo A, Rasovic O, Grossi C, Alamanni F, et al. Does EuroSCORE II perform better than its original versions? A multicentre validation study. Eur Heart J. 2013; 34:22-9.

92. Barili F, Pacini D, Grossi C, Di Bartolomeo R, Alamanni F, Parolari A. Reliability of new scores in predicting perioperative mortality after mitral valve surgery. J Thorac Cardiovasc Surg. 2014; 147:1008-12.

93. Laurent M, Fournet M, Feit B, Oger E, Donal E, Thebault C, et al. Simple bedside clinical evaluation versus established scores in the estimation of operative risk in valve replacement for severe aortic stenosis. Arch Cardiovasc Dis. 2013; 106:651-60.

94. Takkenberg JJ, Kappetein AP, Steyerberg EW. The role of EuroSCORE II in 21st century cardiac surgery practice. Eur J Cardiothorac Surg. 2013; 43:32-3.

95. Nashef SA, Sharples LD. Editorial comment: Pride without prejudice: EuroSCORE II, the STS score and the high-risk patient subset. Eur J Cardiothorac Surg. 2013; 44:1012.

96. Nashef SA, Sharples LD, Roques F, Lockowandt U. EuroSCORE II and the art and science of risk modelling. Eur J Cardiothorac Surg. 2013; 43:695-6.

97. Silva J CM, Reguillo F; Cobiella J, Villagrán E, Montes L, Garcés Z, Ayaon A, Maroto L, Alswies A, Rodríguez E. Validación del EuroSCORE II: ¿funciona en nuestro medio? Cir Cardiov. 2013; 20:59-64.

98. Carnero-Alcazar M, Silva Guisasola JA, Reguillo Lacruz FJ, Maroto Castellanos LC, Cobiella Carnicer J, Villagran Medinilla E, et al. Validation of EuroSCORE II on a single-centre 3800 patient cohort. Interact Cardiovasc Thorac Surg. 2013; 16:293-300.

99. García-Valentín A, Bernabeu E, Pereda D, Josa M, J.M. C, Mestres CA, et al. Validación de EuroSCORE II en España. Cir Cardiov. 2014; 21:246-51.

100. Garcia-Valentin A, Mestres CA, Bernabeu E, Bahamonde JA, Martin I, Rueda C, et al. Validation and quality measurements for EuroSCORE and EuroSCORE II in the Spanish cardiac surgical population: a prospective, multicentre study. Eur J Cardiothorac Surg. 2016; 49:399-405.

101. Clark RE. Outcome as a function of annual coronary artery bypass graft volume. The Ad Hoc Committee on Cardiac Surgery Credentialing of The Society of Thoracic Surgeons. Ann Thorac Surg. 1996; 61:21-6.

102. Birkmeyer JD, Siewers AE, Finlayson EV, Stukel TA, Lucas FL, Batista I, et al. Hospital volume and surgical mortality in the United States. N Engl J Med. 2002; 346:1128-37.

103. Dudley RA, Johansen KL, Brand R, Rennie DJ, Milstein A. Selective referral to high-volume hospitals: estimating potentially avoidable deaths. JAMA. 2000; 283:1159-66.

104. Kansy A, Ebels T, Schreiber C, Tobota Z, Maruszewski B. Association of center volume with outcomes: analysis of verified data of European Association for Cardio-Thoracic Surgery Congenital Database. Ann Thorac Surg. 2014; 98:2159-64.

105. Papadimos TJ, Habib RH, Zacharias A, Schwann TA, Riordan CJ, Durham SJ, et al. Early efficacy of CABG care delivery in a low procedure-volume community hospital: operative and midterm results. BMC Surg. 2005; 5:10.

106. Marcin JP, Li Z, Kravitz RL, Dai JJ, Rocke DM, Romano PS. The CABG surgery volume-outcome relationship: temporal trends and selection effects in California, 1998-2004. Health Serv Res. 2008; 43:174-92.

107. Hannan EL, Kilburn H, Jr., O'Donnell JF, Lukacik G, Shields EP. Adult open heart surgery in New York State. An analysis of risk factors and hospital mortality rates. JAMA. 1990; 264:2768-74.

108. Carey JS, Parker JP, Brandeau C, Li Z. The "occasional open heart surgeon" revisited. J Thorac Cardiovasc Surg. 2008; 135:1254-60.

109. Zacharias A, Schwann TA, Riordan CJ, Durham SJ, Shah A, Papadimos TJ, et al. Is hospital procedure volume a reliable marker of quality for coronary artery bypass surgery? A comparison of risk and propensity adjusted operative and midterm outcomes. Ann Thorac Surg. 2005; 79:1961-9.

110. Shahian DM, O'Brien SM, Normand SL, Peterson ED, Edwards FH. Association of hospital coronary artery bypass volume with processes of care, mortality, morbidity, and the Society of Thoracic Surgeons composite quality score. J Thorac Cardiovasc Surg. 2010; 139:273-82.

111. Flood AB, Scott WR, Ewy W. Does practice make perfect? Part I: The relation between hospital volume and outcomes for selected diagnostic categories. Med Care. 1984; 22:98-114.

112. Luft HS, Bunker JP, Enthoven AC. Should operations be regionalized? The empirical relation between surgical volume and mortality. N Engl J Med. 1979; 301:1364-9.

113. Kouchoukos N, Blackstone E, Doty D, Hanley F, Karp R. Kirklin/Barratt-Boyes Cardiac Surgery. 3 ed. Philadelphia: Churchill Livingstone; 2003.

114. Rutkow IM. Epidemiologic, economic, and sociologic aspects of hernia surgery in the United States in the 1990s. Surg Clin North Am. 1998;78:941-51.

115. Spirou Y, Petrou A, Christoforides C, Felekouras E. History of biliary surgery. World J Surg. 2013; 37:1006-12.

116. Higgins TL, Estafanous FG, Loop FD, Beck GJ, Blum JM, Paranandi L. Stratification of morbidity and mortality outcome by preoperative risk factors in coronary artery bypass patients. A clinical severity score. JAMA. 1992; 267:2344-8.

117. Tu JV, Jaglal SB, Naylor CD. Multicenter validation of a risk index for mortality, intensive care unit stay, and overall hospital length of stay after cardiac surgery. Steering Committee of the Provincial Adult Cardiac Care Network of Ontario. Circulation. 1995; 91:677-84.

118. Huijskes RV, Rosseel PM, Tijssen JG. Outcome prediction in coronary artery bypass grafting and valve surgery in the Netherlands: development of the Amphiascore and its comparison with the Euroscore. Eur J Cardiothorac Surg. 2003; 24:741-9.

119. Pons JM, Granados A, Espinas JA, Borras JM, Martin I, Moreno V. Assessing open heart surgery mortality in Catalonia (Spain) through a predictive risk model. Eur J Cardiothorac Surg. 1997; 11:415-23.

120. The Society of Thoracic Surgeons. Online STS Risk Calculator 2012. Available from: http://riskcalc.sts.org.

121. Josa M, Cortina J, Mestres CA, Pereda D, Walton P, Kinsman R. Primer informe del proyecto español de calidad de cirugía cardiovascular del adulto. Barcelona: Publicaciones Permanyer; 2013.

122. Unger F. Open heart surgery in Europe 1993. Eur J Cardiothorac Surg. 1996; 10(2):120-8.

123. Bustamante-Munguira J, Centella T, Hornero F. Cirugía Cardiovascular en España en el año 2013. Registro de intervenciones de la Sociedad Española de Cirugía Torácica-Cardiovascular. Cir Cardiov. 2014; 21:271-85.

124. Igual A, Mestres CA. Cirugía cardiovascular en España en los años 2006-2008. Registro de intervenciones de la Sociedad Española de Cirugía Torácica-Cardiovascular (SECTCV). Cir Cardiov. 2010; 17:67-83.

125. Goldstone T. EuroSCORE official website: http://www.euroscore.org; 2011.

126. Garcia-Valentin A, Mestres CA. Reply to Nezic. Eur J Cardiothorac Surg. 2016; 49:1021-2.

127. Nezic D. The EuroSCORE II performances in the Spanish cardiac surgical population. Eur J Cardiothorac Surg. 2016; 49:1021.

128. Bridgewater B, Gummert J, Kinsman R, Walton P. Fourth EACTS Adult Cardiac Surgical Database Report. 1st ed. Oxfordshire: Dendrite Clinical Systems LTD; 2010.

129. Guida P, Mastro F, Scrascia G, Whitlock R, Paparella D. Performance of the European System for Cardiac Operative Risk Evaluation II: a meta-analysis of 22 studies involving 145,592 cardiac surgery procedures. J Thorac Cardiovasc Surg. 2014; 148:3049-57.

130. Koszta G, Sira G, Szatmari K, Farkas E, Szerafin T, Fulesdi B. Performance of EuroSCORE II in Hungary: a single-centre validation study. Heart Lung Circ. 2014; 23:1041-50.

131. Siregar S, Groenwold RH, de Heer F, Bots ML, van der Graaf Y, van Herwerden LA. Performance of the original EuroSCORE. Eur J Cardiothorac Surg. 2012; 41:746-54.

132. Geissler HJ, Holzl P, Marohl S, Kuhn-Regnier F, Mehlhorn U, Sudkamp M, et al. Risk stratification in heart surgery: comparison of six score systems. Eur J Cardiothorac Surg. 2000; 17:400-6.

133. Kain ZN. The legend of the P value. Anesth Analg. 2005; 101:1454-6.

134. Grant SW, Hickey GL, Dimarakis I, Trivedi U, Bryan A, Treasure T, et al. How does EuroSCORE II perform in UK cardiac surgery; an analysis of 23 740 patients from the Society for Cardiothoracic Surgery in Great Britain and Ireland National Database. Heart. 2012; 98:1568-72.

135. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. Biometrics. 1987; 43:487-98.

136. Howell NJ, Head SJ, Freemantle N, van der Meulen TA, Senanayake E, Menon A, et al. The new EuroSCORE II does not improve prediction of mortality in high-risk patients undergoing cardiac surgery: a collaborative analysis of two European centres. Eur J Cardiothorac Surg. 2013; 44:1006-11.

137. Kuwaki K, Inaba H, Yamamoto T, Dohi S, Matsumura T, Morita T, et al. Performance of the EuroSCORE II and the Society of Thoracic Surgeons Score in patients undergoing aortic valve replacement for aortic stenosis. J Cardiovasc Surg (Torino). 2015; 56:455-62.

138. Silva J. Validación de EuroSCORE II en España. ¿Y ahora qué hacemos? Cir Cardiov. 2014; 21:237-8.

139. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol. 2008; 61:1085-94.

140. Borrás F, Ferrandis E, Sánchez A, Segura J. Cuadernos de bioestadística II. 1ª ed. Alicante: Universidad de Alicante; 1995.