INVESTIGACIÓN CLÍNICO-EPIDEMIOLÓGICA EN PRONÓSTICO:

DESARROLLOS METODOLÓGICOS Y APLICACIÓN CLÍNICA

UNIVERSIDAD AUTONOMA DE MADRID

Facultad de Medicina

Departamento de Medicina Preventiva y Salud Pública y Microbiología

Borja Manuel Fernández Félix

2021

UNIVERSIDAD AUTÓNOMA DE MADRID

FACULTAD DE MEDICINA

DEPARTAMENTO DE MEDICINA PREVENTIVA Y SALUD PÚBLICA Y MICROBIOLOGÍA

INVESTIGACIÓN CLÍNICO-EPIDEMIOLÓGICA EN PRONÓSTICO: DESARROLLOS METODOLÓGICOS Y APLICACIÓN CLÍNICA

TESIS DOCTORAL

Borja Manuel Fernández Félix

Directores:
Dr. Javier Zamora Romero
Dra. Esther García García-Esquinas

Madrid, 2021

INVESTIGACIÓN CLÍNICO-EPIDEMIOLÓGICA EN PRONÓSTICO: DESARROLLOS METODOLÓGICOS Y APLICACIÓN CLÍNICA

AUTOR: *BORJA MANUEL FERNÁNDEZ FÉLIX*

DIRECTORES: *JAVIER ZAMORA ROMERO Y ESTHER GARCÍA GARCÍA-ESQUINAS*

RESUMEN

En esta Tesis se presentan dos modelos pronósticos para predecir desenlaces importantes en dos contextos clínicos distintos. Ambos modelos han sido desarrollados siguiendo alternativas metodológicas complementarias.

En el primero de ellos, ante la ausencia en la literatura científica de herramientas que permitieran predecir el riesgo de crisis epilépticas durante el embarazo en mujeres en tratamiento antiepiléptico, se desarrolló un modelo pronóstico *de novo* a partir de los datos primarios recogidos en el estudio EMPiRE.

En el segundo contexto clínico, la situación es distinta y se disponía ya de varios modelos en la literatura para predecir el riesgo de mortalidad postoperatoria en pacientes con endocarditis infecciosa. Por ello se optó por hacer una revisión sistemática y meta-análisis para identificar y evaluar la calidad de los modelos disponibles y, a partir de ellos, crear un modelo único (meta-modelo) que fue optimizado para el registro nacional GAMES de endocarditis infecciosa.

En ambos escenarios clínicos, se ha pretendido adicionalmente promover la utilización de estos modelos en la práctica clínica. Para este fin se han creado dos calculadoras online de libre acceso que han sido implementadas en la plataforma web Evidencio (https://www.evidencio.com/). Esta implementación facilita enormemente la obtención de una predicción personalizada del riesgo de los desenlaces considerados dadas las características individuales de los pacientes.

A pesar de que los métodos para el desarrollo y validación de los modelos pronóstico han sido descritos por múltiples autores, estos métodos son en ocasiones complejos para investigadores con conocimientos limitados de estadística. Por ello, es recomendable que el equipo investigador de un estudio de desarrollo o validación de un modelo pronóstico cuente entre sus miembros con algún estadístico o persona con amplio bagaje metodológico. Sin embargo, no siempre es así, por lo que facilitar a los investigadores herramientas útiles y de manejo sencillo, como el comando – `bsvalidation` – para el software Stata que se ha presentado en el tercer estudio de la tesis, puede ayudar a paliar los efectos de las carencias metodológicas del equipo investigador.

ABSTRACT

In this thesis, two prognostic models are presented to predict important outcomes in two different clinical contexts. Both models have been developed following complementary methodological alternatives.

In the first of these, given the absence in the scientific literature of tools that would allow predicting the risk of epileptic seizures during pregnancy in women on antiepileptic treatment, a new prognostic model was developed from the primary data collected in the EMPiRE study.

In the second clinical context, the situation is different and several models were already available in the literature to predict the risk of postoperative mortality in patients with infective endocarditis. For this reason, it was decided to carry out a systematic review and meta-analysis to identify and evaluate the quality of the available models and, based on them, create a single model (meta-model) that was optimized for the GAMES national registry of infective endocarditis.

In both clinical settings, it has been further intended to promote the use of these models in clinical practice. For this purpose, two free access online calculators have been created that have been implemented on the Evidencio web platform (https://www.evidencio.com/). This implementation greatly facilitates obtaining a personalized prediction of the risk of the outcomes considered given the individual characteristics of the patients.

Although the methods for the development and validation of prognostic models have been described by multiple authors, these methods are sometimes complex for researchers with limited knowledge of statistics. For this reason, it is recommended that the research team of a study of development or validation of a prognostic model has among its members a statistician or person with extensive methodological background. However, this is not always the case, so providing researchers with useful and easy-to-use tools, such as the – `bsvalidation` – command for Stata software that has been presented in the third study of the thesis, can help alleviate the effects of the methodological deficiencies of the research team.

A la memoria cariñosa de mi madre

Ana, artífice de lo que hoy soy

y quien sigue guiando mis pasos…

… y a las que en su ausencia se comportaron y quise como madres

Toñi, la abuela Aurora y la abuela Pepa,

y que se fueron mientras escribía estas líneas.

A todas GRACIAS.

"Para maximizar el beneficio para la sociedad,
no sólo es necesario investigar sino hacerlo bien."

Profesor Douglas Altman

AGRADECIMIENTOS

Me gustaría aprovechar este espacio para agradecer a tantas personas que han ayudado a cumplir mis metas.

En primer lugar, a Mar, porque sin haber escrito una línea de esta tesis puedes sentirte coautora, pues sin ti no lo hubiera logrado, por tu interés y curiosidad, y por tu apoyo incondicional incluso en tus peores momentos.

A mis directores de tesis, la Dra. Esther García, por su apoyo constante, consejos y enseñanzas durante todo este periodo, y el Dr. Javier Zamora, por todo ello y por haber confiado en mí desde los inicios de mi carrera investigadora, a los dos os agradezco vuestros "Borja, está genial", pero te devuelvo el documento teñido de rojo con varios cientos de modificaciones, y digo yo "jod…, pues para estar genial…."

A Alfonso Muriel, mi compañero y amigo, por su insistencia en que me involucrase en este proyecto, por sus pragmáticos consejos y sus "te quieres poner con la tesis y dejarte de rollos". Junto a Javier mis maestros en lo profesional y ejemplos en lo personal.

Al resto de compañeros que están o han estado acompañando mi trayectoria en la Unidad de Bioestadística Clínica, y de los que he aprendido y sigo aprendiendo cada día, Víctor, Nieves, Ana, David, Jesús, Elena, Ingrid y Andrea entre otros.

A Shakila Thangaratinam y John Allotey, compañeros de la Universidad de Birmingham, por su generosidad y por permitirme participar en estudios de gran relevancia clínica e interés metodológico.

A Laura Varela, Cirujana del Hospital Fundación Jiménez Díaz, por su disposición siempre conmigo y por darme la oportunidad de aplicar los métodos empleados en esta tesis en una pregunta de investigación relevante.

Al Grupo de Apoyo al Manejo de la Endocarditis infecciosa en España (GAMES), por su generosa cesión de los datos del registro nacional para poder realizar parte de los análisis estadísticos de esta tesis.

A las mujeres reclutadas en el estudio EMPiRE y los pacientes del registro GAMES por consentir a que se empleen sus datos clínicos para realizar investigaciones científicas cuyos resultados beneficiarán a otros pacientes.

A todos los coautores y compañeros que han participado en los estudios incluidos en esta tesis, porque cada uno de vosotros ha aportado su conocimiento y tiempo para mejorar la investigación del pronóstico.

Y por supuesto, a todos y todas aquellas que siempre habéis confiado en mí y me habéis ayudado a crecer como persona, mi hermano Alex, mi padre Manolo, la tata Toñi, las tías Pilar y Montse, primos, primas, tíos, tías, cuñados, cuñadas, amigos y amigas que os alegráis con mis alegrías.

ÍNDICE

ÍNDICE

Introducción

1. INTRODUCCIÓN

1.1. INVESTIGACIÓN DEL PRONÓSTICO

Pronosticar, según la Real Academia Española (RAE) de la lengua, significa predecir algo futuro a partir de indicios. En investigación clínico-epidemiológica el pronóstico lo identificamos con la probabilidad o el riesgo de que un individuo desarrolle un particular estado de salud en un tiempo determinado basado en sus características (1). El pronóstico, junto con el diagnóstico de la enfermedad o condición de salud, y el tratamiento o cuidado del individuo, forma parte de la tríada de la práctica clínica.

En la época hipocrática el pronóstico era el elemento más importante de los tres, pues se conocía poco de los elementos diagnósticos de las enfermedades, y el tratamiento en muchos casos se limitaba al conocido – *primum non nocere* –. Durante esta época el pronóstico se focalizaba en el probable curso natural de la enfermedad basado en las observaciones previas de los pacientes con una condición similar (2). El interés en un diagnóstico preciso y el aumento de las opciones en el tratamiento fue dejando de lado el interés por el pronóstico. En el siglo XX, a pesar de que el diagnóstico se mantenía como la actividad clínica más importante, el pronóstico fue recuperando interés (3). Hoy día, los avances en el conocimiento de los mecanismos de las enfermedades, así como el creciente interés por la biología molecular han vuelto a situar la investigación del pronóstico como *trending topic*.

La información del pronóstico no sólo es importante para los clínicos y sus pacientes durante la práctica clínica diaria y en el proceso de toma de decisiones, sino también para las instituciones sanitarias y políticas, y para los individuos con una determinada condición de salud tal como un embarazo.

La investigación del pronóstico tiene como objetivos resumir, explicar y predecir desenlaces – *outcomes* – futuros relevantes, proporcionando información sobre la salud y el bienestar futuro de personas con enfermedades o ciertas condiciones de salud.

Son miles los estudios que se publican anualmente para responder preguntas de investigación sobre pronóstico, aunque los términos referentes a este tipo de estudios eran tradicionalmente confusos.

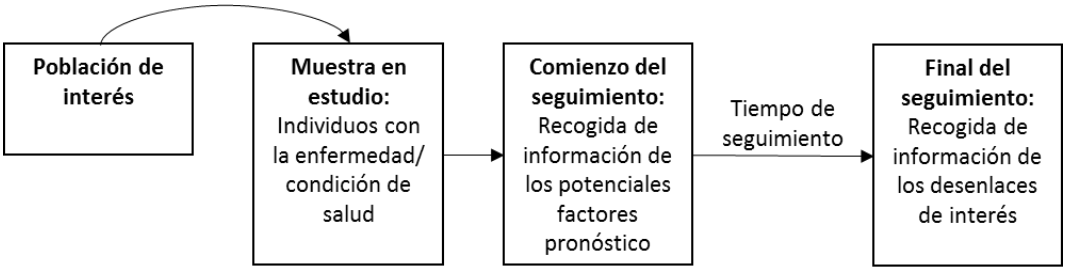


FIGURA 1. ESTRUCTURA DE UN ESTUDIO DE INVESTIGACIÓN DEL PRONÓSTICO

En 2013, la asociación *PROGnosis RESearch Strategy* (PROGRESS) publicó una serie de artículos presentando la estructura de los cuatro tipos de estudio en investigación en pronóstico (4–7):

- **PROGRESS tipo I:** Estudios de pronóstico general. En estos estudios se analizan los resultados en muestras de pacientes con una determinada enfermedad o situación de salud de interés. Este tipo de estudios evalúan los desenlaces de forma global, es decir, responden a cuál es el valor promedio, o el riesgo, del desenlace de interés.

- **PROGRESS tipo II:** Estudios de factores pronósticos. En estos estudios se evalúan qué características (factores) están asociadas con cambios en los resultados globales para los individuos del estudio. Un factor pronóstico es una variable asociada con el riesgo de un desenlace concreto en individuos con una determinada condición de salud. Así, diferentes valores o categorías de un factor pronóstico se asocian con diferentes resultados pronósticos.

- **PROGRESS tipo III:** Estudios de modelos pronósticos. Estos estudios se centran en el desarrollo, validación y evaluación del impacto de modelos que incorporan múltiples factores pronósticos para predecir a nivel individual el valor, o el riesgo, del desenlace de interés.

- **PROGRESS tipo IV:** Estudios de predictores del efecto de un tratamiento. Estudian las características que predicen si un individuo responderá o no a un determinado tratamiento.

Dado que el foco de esta tesis se relaciona con los estudios de tipo III, a continuación se presentan ciertas generalidades a cerca de los modelos pronósticos.

1.1.1. MODELO PRONÓSTICO

Debido a la variabilidad existente entre individuos y enfermedades (cuya etiología, presentación, y tratamiento es con frecuencia variable), la información de un único factor pronóstico raramente es suficiente para dar una estimación adecuada de la predicción de un desenlace de interés. Por ello, para realizar predicciones individualizadas se requiere de la combinación de múltiples factores pronósticos, también llamados predictores, haciendo necesario el uso de métodos de análisis multivariables. Estas herramientas son comúnmente llamadas modelos pronósticos o predictivos, reglas de predicción, puntuaciones o scores de riesgo. Los estudios que abordan el desarrollo de tales herramientas se clasifican como PROGRESS tipo III.

Los modelos pronósticos son habitualmente desarrollados usando métodos de regresión multivariable. Estos métodos proporcionan una ecuación matemática para estimar a nivel individual el valor esperado de un desenlace continuo, el riesgo de un desenlace binario o el tiempo hasta el evento de interés. En el caso de desenlaces binarios como los evaluados en la presente tesis doctoral, la ecuación vendría definida por:

$$ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

ECUACIÓN 1. ECUACIÓN DE UN MODELO DE REGRESIÓN LOGÍSTICA

Donde, p_i es la probabilidad de desarrollar el desenlace por el individuo i-ésimo, que dependerá de la constante del modelo (α), del valor de los predictores incluidos en el modelo ($X_{1i}, X_{2i}, \dots X_{ki}$) y de sus correspondientes coeficientes de regresión ($\beta_1, \beta_2, \dots \beta_k$). Una vez desarrollado un modelo pronóstico, la probabilidad del desenlace puede ser estimada usando la siguiente expresión:

$$p_i = \frac{exp(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})}{1 + exp(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})}$$

ECUACIÓN 2. ESTIMACIÓN DE LA PROBABILIDAD DE DESARROLLAR EL DESENLACE A PARTIR DE UN MODELO DE REGRESIÓN LOGÍSTICA

1.2. INVESTIGACIÓN PRIMARIA DE ESTUDIOS DE MODELOS PRONÓSTICO

Para plantear un estudio de desarrollo de un modelo pronóstico debe existir una necesidad. Es decir, que la predicción del desenlace sea de interés y que no existan otros modelos válidos para predecir el mismo desenlace. Una vez desarrollado, el modelo debe demostrar buen rendimiento predictivo en la validación externa y evidenciar un impacto positivo en los resultados de salud antes de ser implementado en la práctica clínica. Sin embargo, muchos modelos pronósticos desarrollados nunca llegan a ser utilizados en la práctica clínica debido a la baja calidad de los estudios de desarrollo y validación de modelos pronósticos (8), y a las dificultades de implementación que persisten incluso cuando los modelos han sido desarrollados e informados siguiendo los estándares recomendados (9,10). Estas dificultades son explicadas en parte por la falta de interés de los profesionales del campo donde debe ser aplicado, o debido a que, en ocasiones, los usuarios de estas herramientas no saben cómo integrar los resultados del modelo en la toma de decisiones (11). En otras casos, los modelos desarrollados incluyen predictores de difícil accesibilidad o elevado coste, lo que reduce las oportunidades de aplicabilidad (12).

La investigación de modelos pronósticos tiene tres importantes fases: el desarrollo del modelo (incluyendo la validación interna), la validación externa y la evaluación de su impacto (1,13).

1.2.1. DESARROLLO DE UN MODELO PRONÓSTICO

El desarrollo de un nuevo modelo pronóstico usando datos primarios de un estudio epidemiológico es muy costoso. Para evitar desperdicio de recursos de investigación (14), el desarrollo de un nuevo modelo predictivo debe estar siempre supeditado a la ausencia de modelos previos válidos. Además de esta recomendación preliminar, una condición necesaria obvia para determinar la necesidad de desarrollar un modelo pronóstico fiable a partir de datos primarios de un estudio es que éstos datos sean de calidad. Por último, el desarrollo de un buen modelo también requiere de la aplicación de la metodología adecuada en cada paso, desde el diseño del protocolo de investigación hasta la presentación y comunicación del modelo pronóstico y la evaluación de su rendimiento predictivo. Si ya existen modelos pronósticos que permiten estimar de forma precisa la probabilidad del desenlace de interés, no es conveniente desarrollar un nuevo modelo. En tales situaciones, los investigadores deben diseñar estudios de validación que permitan evaluar el rendimiento predictivo de los modelos existentes cuando son aplicados a otros sujetos (15–18).

Los datos que se emplean para el desarrollo de un modelo pronóstico deben ser de alta calidad y, en este sentido, el mejor diseño para obtenerlos son los estudios de cohortes prospectivos (19). El diseño prospectivo permite minimizar los potenciales sesgos de selección e información al establecer claramente los criterios de inclusión y exclusión de pacientes en el estudio, y al establecer las definiciones y métodos de medición tanto de los desenlaces como de los predictores. Este diseño permite también reducir el número de datos faltantes. Sin embargo, diseñar una cohorte prospectiva es algo muy costoso económicamente, e implica un periodo de seguimiento habitualmente largo para permitir la observación de un número relevante de desenlaces. Por ello, son útiles para los investigadores los datos existentes en registros retrospectivos, en ensayos clínicos o en bases de datos electrónicas de registros clínicos.

El tamaño de la muestra es importante para establecer estrategias de selección de predictores más fiables y obtener coeficientes de regresión más precisos (20). En regresión logística el tamaño de muestra efectivo es el mínimo entre el número de eventos (individuos con el desenlace de interés) y no eventos (individuos sin el desenlace de interés), y éste limitará el número de candidatos predictores que pueden ser evaluados en el análisis. La regla de oro que muchos investigadores utilizan es de que se necesitan al menos diez eventos por predictor (EPP) evaluado en el desarrollo del modelo (21). Sin embargo, esta regla resulta demasiado simple (22–24) y, recientemente, se han publicado varias herramientas que permiten calcular el tamaño de muestra adecuado en base también a otros parámetros tales como la prevalencia del desenlace de interés y las medidas del rendimiento y ajuste esperado (25).

Un buen modelo pronóstico debe demostrar un buen rendimiento predictivo en individuos que no han sido utilizados en el desarrollo del modelo (validación externa). Cuando el rendimiento del modelo es evaluado en los mismos sujetos a partir de los cuales fue desarrollado se obtiene el conocido rendimiento aparente. En esta situación, las medidas de rendimiento tienden a ser demasiado optimistas indicando que el modelo funciona mejor de lo que realmente lo hará cuando sea aplicado a otros individuos nuevos (26). Para evaluar el rendimiento de un modelo de regresión logística se debe resumir el ajuste global, la calibración y la discriminación (27–29).

- **Ajuste global:** Los estadísticos de ajuste global miden el rendimiento general del modelo. Los estadísticos R^2 de *Cox-Snell* (30) o el R^2 de *Nagelkerke* (31) son alguna de las alternativas al coeficiente de determinación (R^2), frecuentemente, utilizado como medida del ajuste global en regresión lineal. Otra medida del ajuste global es el error cuadrático medio de las predicciones, denominado *Brier score* (32).

- **Calibración:** Refleja el grado de acuerdo entre las predicciones estimadas por el modelo y los resultados observados. Las medidas más adecuadas para resumir la calibración de un modelo logístico son: el cociente entre eventos esperados y eventos observados (en inglés: *E/O ratio*), que resume la calibración global; la *Calibration-in-the-Large* (CITL) (no existe una traducción al español de modo que se empleará el término en inglés en el documento de la tesis), muy relacionada con el *E/O ratio*, y que compara la media de los riesgos predichos por el modelo con la media de los riesgos observados e indica en qué medida las predicciones del modelo son sistemáticamente demasiado bajas o altas; y la pendiente de calibración (en inglés: *calibration slope*), que mide el grado de sobreajuste del modelo (19,28,29). Desafortunadamente, la prueba estadística más utilizada por los investigadores es test estadístico de *Hosmer-Lemeshow* (33) que resulta insuficiente para valorar la calibración. Su uso es desaconsejado porque el test no indica la dirección de la mala calibración, ni el p-valor refleja el grado de ausencia de calibración, pues es una combinación de la falta de calibración, del tamaño de la muestra y de la agrupación de los datos (9,10,28,29). La calibración puede ser resumida a través de gráficos. Un gráfico de calibración enfrenta las probabilidades predichas por el modelo (en el eje X) frente a las probabilidades observadas (en el eje Y). Se pueden agrupar los sujetos de estudio por deciles de riesgo y/o aplicando funciones de suavizado (34). Con este gráfico de calibración podemos ver la dirección y la magnitud de una mala calibración, si la hubiera, a través del rango de probabilidades predichas.

- **Discriminación:** Indica la capacidad del modelo para distinguir entre los individuos que experimentan el desenlace de interés y los que no. Un modelo predictivo discrimina perfectamente cuando la probabilidad predicha para todos los individuos que tienen el evento de interés es mayor que la de todos los individuos que no lo tienen. La capacidad discriminante de un modelo de regresión logística frecuentemente se evalúa usando el estadístico de concordancia *C* (20). Este refleja la probabilidad de que seleccionados al azar un par de individuos, uno con el evento y otro sin el evento, el modelo asigne la probabilidad más alta al individuo con el evento (35).

En la sección de resultados, en el apartado de justificación y en la discusión de los aspectos metodológicos del estudio 3 de esta tesis se presentan con más detalle las medidas del rendimiento predictivo de los modelos de regresión logística.

El rendimiento predictivo del modelo y el método utilizado para la su estimación debe ser informado en los resultados de un estudio de desarrollo de un modelo pronóstico. La declaración TRIPOD (*Transparent Reporting of multivariable prediction model for Individual Prognosis or Diagnosis*) es un conjunto de recomendaciones para comunicar de forma completa y estandarizada los estudios de modelos de predicción en ciencias biomédicas (9,10). Incluye 22 ítems que deben ser reportados con el objetivo de mejorar la comunicación transparente de los estudios de desarrollo, validación o actualización de modelos de predicción, ya sean éstos desarrollados con fines diagnósticos o pronósticos.

1.2.2. VALIDACIÓN INTERNA DE UN MODELO PRONÓSTICO

El objetivo de la validación interna es reportar de forma fiable cual será el rendimiento predictivo del modelo cuando se aplique en otros individuos (17). La validación interna forma parte del proceso de desarrollo del modelo pronóstico. Es esencial realizar este proceso para evaluar la posible existencia de sobreajuste en el modelo, en particular, cuando el tamaño de la muestra utilizada es demasiado pequeño para la cantidad de candidatos predictores que han sido evaluados. Un modelo sobreajustado se adapta muy bien a los datos en los que ha sido desarrollado exagerando su capacidad predictiva. Sin embargo, cuando el modelo es aplicado a otros datos su rendimiento decrece (26,36,37). En caso de que el modelo ajustado presente cierto grado de sobreajuste es necesario corregirlo. Una óptima validación interna considera el potencial sobreajuste en el modelo para reportar unas medidas del rendimiento predictivo más realistas corregidas por el exceso de optimismo.

A pesar de la importancia de la validación interna, este es uno de los principales déficits de los estudios de desarrollo de modelos pronósticos. Es frecuente que los modelos no sean validados internamente, y cuando lo son, no lo sean de forma adecuada (8,38,39).

Entre la lista de ítems que la declaración TRIPOD recomienda comunicar en los estudios de desarrollo de un modelo predictivo, en el apartado de métodos de análisis estadísticos se incluye un ítem para especificar el método de validación interna empleado en el desarrollo del modelo (9,10).

Existen diferentes técnicas de validación interna:

- **Validación del rendimiento aparente.** Consiste en evaluar el rendimiento predictivo del modelo utilizando los mismos datos empleados en el desarrollo del mismo. Desafortunadamente, son muchos los modelos en los que solo se comunica el

rendimiento predictivo aparente. En la validación aparente el modelo suele mostrar un rendimiento excesivamente optimista. Solo en modelos desarrollados a partir de tamaños de muestra enormes debería ser empleado como único método de validación (20).

- **Validación por división de los datos (*splitting*).** Consiste en dividir la muestra original en dos submuestras, una donde se desarrolla el modelo – *training* – y otra donde se valida – *test* –. Este método de validación es ineficiente debido a que no utiliza todos los datos disponibles para el desarrollo del modelo, y las diferentes divisiones que se hagan de la muestra original conducirán a distintos resultados (40). La forma más habitual de división de la muestra es hacerlo al azar, aunque esto está desaconsejado. Hay varias alternativas para realizar esta división de la muestra original evitando que sea el azar el que divida los subconjuntos de desarrollo y validación del modelo pues este azar hace que ambos grupos sean excesivamente similares. Cuando el tamaño de la muestra es suficientemente amplio se ha propuesto dividir la muestra de acuerdo al tiempo (validación temporal) o de acuerdo a la localización de reclutamiento (validación geográfica) (18).

- **Validación cruzada.** Es una extensión de las técnicas de *splitting* después de desarrollar el modelo usando todos los datos disponibles. Consiste en dividir o segmentar el conjunto de datos en *k* grupos de igual (o similar) tamaño. Se utilizan *k-1* subconjuntos como *training* y el subconjunto restante se utiliza como *test*. Se repite este procedimiento *k* veces, hasta que cada uno de los subconjuntos ha sido utilizado una vez como grupo de validación. Finalmente, obtenemos la medida de rendimiento como el promedio de las medidas calculadas sobre las *k* muestras (41).

- **Validación bootstrapping.** Este método utiliza para la validación todos los datos empleados en el desarrollo del modelo permitiendo, además, cuantificar el exceso de optimismo en las medidas de rendimiento del modelo. En la siguiente sección se desglosa con detalle esta metodología.

VALIDACIÓN INTERNA BOOTSTRAPPING

La validación interna mediante técnicas de remuestreo (bootstrapping) es considerada la metodología más adecuada para obtener una estimación más fiable del rendimiento predictivo del modelo. El término bootstrap parece proceder de una antigua leyenda alemana sobre el barón Münchhausen. Esta leyenda narra las increíbles historias de aventuras que el barón, algo fanfarrón, contaba a sus conciudadanos a la vuelta de sus viajes. Entre ellas decía que fue capaz de salir de un pantano (junto a su caballo) levantándose de su propio cabello (ilustración 1). Aunque posteriores versiones de la leyenda dicen que este usó las correas de sus botas (traducción al inglés: *boot straps*) para salir del agua, dando lugar al término "bootstrapping", que hace referencia a algo que se consigue por tus propios medios.

ILUSTRACIÓN 1. ILUSTRACIÓN DEL BARÓN MÜNCHHAUSEN BY MARTIN DISTELI, PUBLIC DOMAIN, VIA WIKIMEDIA COMMONS

En el ámbito de la estadística, el bootstrapping hace referencia al método empleado para estimar la distribución muestral de un estimador extrayendo muestras con reemplazamiento de la muestra original, conocidas como muestras bootstrap (42).

La figura 2 presenta el proceso de muestreo mediante técnicas bootstrapping. En la muestra original se dispone de 10 individuos distintos, representados por diferentes tramas de relleno. Las subsiguientes muestras (muestras bootstrap) son del mismo tamaño de la muestra original (10 individuos) que han sido seleccionados desde la muestra original con reemplazamiento. De

modo que por ejemplo en la muestra bootstrap 1, el individuo con relleno de líneas diagonales es seleccionado tres veces, los individuos con relleno sólido blanco y sólido negro son incluidos dos veces cada uno, y el individuo con relleno sólido gris no ha sido seleccionado. En cambio para la muestra bootstrap 2, el individuo con relleno de líneas diagonales no es seleccionado, el individuo con relleno sólido negro de nuevo es seleccionado dos veces, y los individuos con relleno sólido blanco y con relleno sólido gris ahora aparecen una vez cada uno. Al final del proceso de remuestreo se dispone de *b* muestras bootstrap obtenidas desde la original.

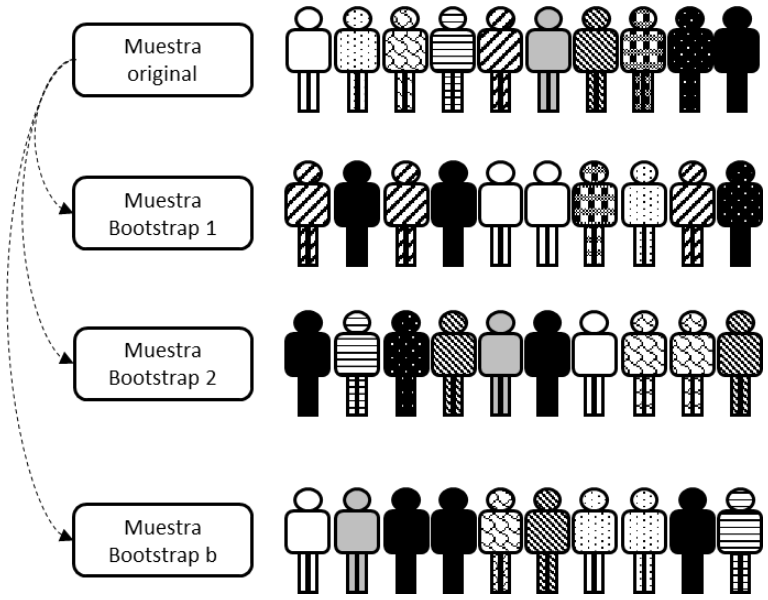


FIGURA 2. REMUESTREO MEDIANTE TÉCNICAS BOOTSTRAPPING.

Uno de los principales usos de las técnicas de bootstrapping se tiene en la validación interna de modelos de regresión (43). La aplicación del proceso se muestra a continuación:

En primer lugar, el modelo pronóstico se desarrolla en la muestra original de datos (paso 1) y se estima el rendimiento aparente (paso 2). Una vez estimado el modelo, desde la muestra original se obtienen *b* muestras bootstrap (por ejemplo, *b* = 1.000) (paso 3), y en cada una de ellas se repite el mismo proceso utilizado para el desarrollo del modelo en la muestra original. De este modo, de cada muestra bootstrap se obtiene un modelo distinto (modelos bootstrap) (paso 4). Cada modelo bootstrap es evaluado en su muestra correspondiente para obtener el rendimiento aparente bootstrap (paso 5), a la vez que en la muestra original para obtener el test de rendimiento (paso 6). Entonces, el exceso de optimismo del modelo se calcula como el promedio

de las diferencias entre el rendimiento aparente de cada muestra bootstrap y el test de rendimiento en la muestra original.

Finalmente, se puede obtener el rendimiento del modelo corregido por el optimismo, simplemente, restando el exceso de optimismo al rendimiento aparente en la muestra original. La figura 3 muestra el esquema del proceso de validación interna bootstrap.

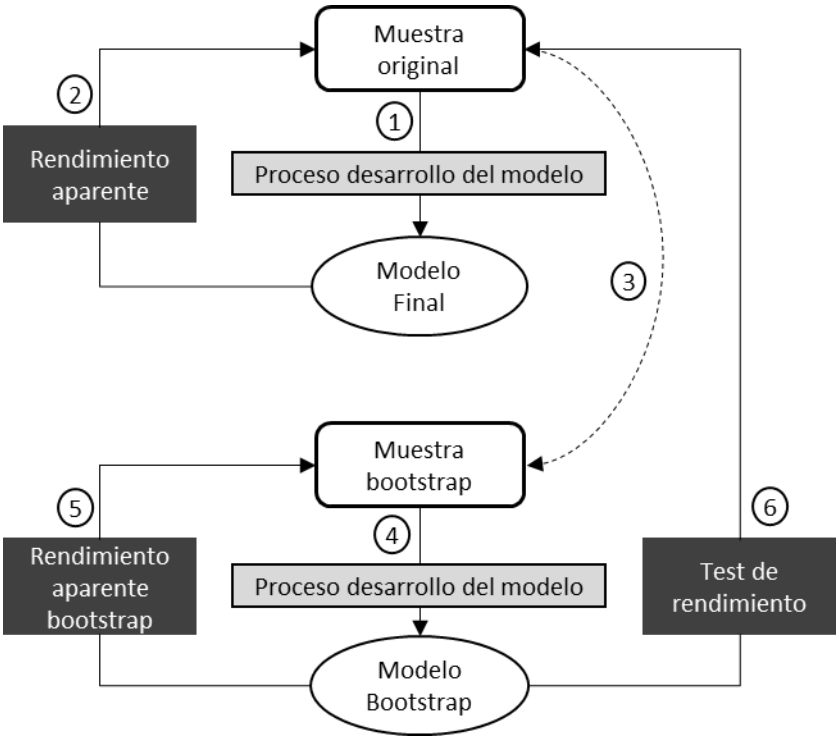


FIGURA 3. ESQUEMA DE VALIDACIÓN INTERNA BOOTSTRAP

El sobreajuste es un problema frecuente en el desarrollo de modelos pronósticos y su presencia amenaza la validez de las predicciones que se obtengan. Este problema se traduce en que se obtienen estimaciones del rendimiento del modelo demasiado optimistas y predicciones demasiado extremas. El sobreajuste se produce cuando el modelo es demasiado complejo para la cantidad de información disponible. Una alternativa para resolver el problema de sobreajuste consiste en penalizar los coeficientes de regresión del modelo. El factor de penalización o –*shrinkage factor* – puede ser estimado mediante las técnicas de bootstrapping (20). El factor de penalización bootstrap es determinado por el valor de la pendiente de calibración en la validación interna bootstrapping. Otro factor de penalización es el factor heurístico propuesto *Van Houwelingen* y *le Cessie* (44). Ambos métodos aplican un factor de ajuste uniforme para

todos los coeficientes del modelo. Sin embargo, existen otras técnicas más sofisticadas de métodos de regresión que directamente penalizan la función de verosimilitud en lugar de emplear un factor de penalización *post-hoc*. Una alternativa popular es el método LASSO (*Least Absolute Shrinkage and Selection Operator*) el cual emplea un factor de penalización que puede reducir el efecto de un predictor hasta cero, es decir, puede penalizar el efecto del predictor hasta excluirlo del modelo (45). Esta metodología, por tanto, combina el proceso de selección de predictores y la penalización de los coeficientes.

1.2.3. VALIDACIÓN EXTERNA DE UN MODELO PRONÓSTICO

La validación interna evalúa la validez del modelo en el entorno en el que se originaron los datos de desarrollo. Pero un modelo pronóstico debe ser validado en sujetos que no hayan sido empleados en el proceso de desarrollo del modelo. Esto es conocido como validación externa. Para que un modelo pueda ser implementado en la práctica clínica tiene que haber demostrado un buen rendimiento en la validación externa. Sin embargo, varias revisiones sistemáticas en diferentes ámbitos de la medicina, han mostrado que son muy pocos los modelos desarrollados que han sido validados externamente (8,38,39).

En el contexto de validación externa conceptos como la reproducibilidad o transportabilidad son importantes. El término de reproducibilidad hace referencia a la validación externa del modelo en una cohorte similar, en cuanto a población objetivo y ámbito de aplicación, a la utilizada para el desarrollo del modelo. La transportabilidad hace referencia a la validación externa del modelo en distintos ámbitos sanitarios (por ejemplo, un modelo desarrollado en atención especializada debería demostrar buen rendimiento en atención primaria para poder ser implementado de un modo más general), o en diferentes contextos temporales y/o geográficos (por ejemplo, un modelo que es desarrollado en un país de altos ingresos debería demostrar validez en países de bajos o medios ingresos para poder ser implementado en regiones con unas condiciones económicas más desfavorables) (16,19).

Antes de desarrollar un nuevo modelo pronóstico, es preferible que los investigadores validen los modelos existentes en diferentes ámbitos. En caso de un rendimiento deficiente, simples ajustes en el modelo, tales como una adecuada recalibración que permita ajustar el riesgo basal predicho por el modelo a la población de validación (46,47), o una actualización incluyendo nuevos predictores, podrían suponer una mejora aceptable que pudieran hacer mejorar el rendimiento del modelo (48–51).

1.2.4. EVALUACIÓN DEL IMPACTO DE UN MODELO PRONÓSTICO

Los modelos pronósticos pueden influir en los resultados de salud y coste-efectividad de las intervenciones cuando las probabilidades estimadas por el modelo se emplean en el proceso de toma de decisiones clínicas. Esto es así, por ejemplo, cuando empleamos un cierto umbral de probabilidad de un desenlace para decidir si realizar una particular intervención.

Los métodos de análisis de decisión tienen como objetivo evaluar la utilidad clínica de un modelo asignando pesos a cada una de las posibles consecuencias que se derivan de usar las predicciones del modelo en la toma de decisiones. Se trata de identificar las posibles acciones y sus consecuencias y por tanto poder informar la selección de la acción con mejores consecuencias esperadas. Los análisis de curva de decisión permiten evaluar el impacto clínico del modelo utilizando los mismos datos empleados en el desarrollo del modelo (27,52).

1.3. INVESTIGACIÓN DE SÍNTESIS DE ESTUDIOS DE MODELOS PRONÓSTICO

1.3.1. MODELOMANÍA

En la actualidad, en muchas áreas de medicina, el afán de los investigadores por el desarrollo de modelos pronósticos ha desencadenado una amalgama de modelos que han sido desarrollados para una misma población diana y un mismo desenlace de interés. A pesar de la existencia de esta plétora de modelos, los investigadores continúan diseñando nuevos estudios de desarrollo de modelos pronósticos, a menudo utilizando datos de baja calidad y métodos analíticos deficientes. Consecuentemente, muchos de estos modelos presentan carencias importantes en su diseño y análisis, y nunca han sido ni serán utilizados en la práctica clínica. Esta circunstancia enfatiza la importancia de las revisiones sistemáticas de modelos predictivos como herramienta de investigación de síntesis con la que identificar todos los modelos publicados acerca de una misma población y desenlace y evaluar su calidad metodológica y su riesgo de sesgo.

Una reciente revisión sistemática “viva” ha identificado en su última actualización más de 60 modelos predictivos para el diagnóstico y el pronóstico de la COVID-19. Esta revisión concluye que la mayoría de los modelos propuestos se comunican de manera deficiente, tienen un alto riesgo de sesgo y el rendimiento reportado probablemente sea excesivamente optimista. Los autores de esta revisión finalmente no recomiendan ninguno de los modelos de predicción encontrados para que se use en la práctica clínica actual (8). Pero esto no es exclusivo de una enfermedad de reciente aparición como la COVID-19, sino que es común en múltiples campos de la medicina (53,54).

1.3.2. REVISIÓN SISTEMÁTICA

Las decisiones en el contexto sanitario deben tomarse basadas en la síntesis de la mejor evidencia disponible y no sólo en los resultados de un único estudio o basarse en una opinión subjetiva de expertos. En este sentido, en las últimas décadas se ha incrementado notablemente la investigación y el desarrollo de métodos para realizar esta síntesis de la evidencia disponible.

Las revisiones sistemáticas son estudios con una metodología explícita que tienen como objetivo identificar, valorar y resumir todos los estudios que responden a una específica pregunta de investigación (55). Las revisiones sistemáticas son la mejor fuente de evidencia científica y son una herramienta esencial para la toma de decisiones en salud (56).

Las fases para desarrollar una revisión sistemática son comunes para cualquier pregunta de investigación: 1.) formulación de forma estructurada de la pregunta de investigación, 2.) elaboración (y publicación) de un protocolo detallado describiendo los pasos que se van a seguir en la investigación, 3.) búsqueda y selección de los estudios primarios en la literatura, 4.) evaluación de la calidad de los estudios primarios, 5.) síntesis de los hallazgos y 6.) reporte de los resultados y conclusiones (55).

Los métodos de las revisiones sistemáticas de estudios de intervenciones están bien consolidados y protocolizados y son ampliamente utilizados (55). En la investigación del pronóstico, los métodos, manuales y recomendaciones aún están en desarrollo y, por tanto, ocasionan dificultades para los investigadores (57). Pero el creciente interés en el estudio del pronóstico ha llevado a la creación de un grupo específico de la Colaboración Cochrane, el Grupo Cochrane de Métodos de Pronóstico. Este grupo se estableció con el objetivo de desarrollar una metodología sobre cómo se puede mejorar la validez y precisión de las revisiones sistemáticas de estudios de pronóstico (58).

1.3.3. REVISIONES SISTEMÁTICAS DE ESTUDIOS DE MODELOS PRONÓSTICOS

La aplicación de revisiones sistemáticas en la investigación del pronóstico apenas está comenzando, y a pesar de que los principios básicos son similares a los de cualquier otra pregunta de investigación existen numerosos retos que afrontar en todas las fases de la revisión sistemática.

1.3.4. PREGUNTA DE REVISIÓN

La pregunta de investigación en los estudios de pronóstico, tanto de factores como de modelos, debe ser formulada de una forma estructurada y cuidadosa pues los términos de búsqueda y los criterios de inclusión dependerán de ella. En las revisiones sistemáticas de estudios de intervención el acrónimo PICO está ampliamente extendido para establecer la pregunta de investigación en términos de Población, Intervención, Comparador y *Outcome* (desenlace). En una revisión sistemática de pronóstico se debe emplear una versión extendida, el acrónimo PICOTA (en inglés: *PICOTS*) (57):

- Población: población general en la que se desarrolla y aplica el modelo pronóstico.
- Índice: modelo pronóstico que se está analizando (no siempre aplica).
- Comparador(es): Otro modelo pronóstico con el que se desea comparar el índice pronóstico bajo revisión. (no siempre aplica)
- Outcome (o desenlace): resultado o evento que se está interesado en predecir.
- Tiempo: cuándo se miden los predictores y el lapso de tiempo para la predicción del desenlace.
- Ámbito: el escenario donde se utiliza el modelo pronóstico.

1.3.5. PROTOCOLO

El protocolo es un documento en el que se redacta el plan detallado de lo que se va a hacer y cómo se va a llevar a cabo en la revisión sistemática. El protocolo incluye el razonamiento y la justificación de la revisión, los objetivos, los criterios de elegibilidad de los estudios, el método de extracción de datos, la evaluación del riesgo de sesgo, los métodos estadísticos para sintetizar la ”evidencia“, y la comunicación clara y completa de los resultados. El protocolo puede ser registrado en el registro específico de protocolos de Revisiones sistemáticas como PROSPERO que depende de la Universidad de York (59).

1.3.6. BÚSQUEDA Y SELECCIÓN DE ESTUDIOS

Los estudios de modelos pronósticos no están indexados de forma explícita en las bases de datos bibliográficas. Esta situación dificulta una búsqueda eficiente de la literatura, incrementando el número de estudios irrelevantes que deben ser revisados. Recientemente se han desarrollado filtros metodológicos para concretar de forma más precisa los criterios de búsqueda (60). Una combinación adecuada de los filtros y palabras clave relacionadas con la población y el

desenlace, permiten concretar la búsqueda y reducir el número necesario de estudios a leer (en inglés: *number needed to read* (NNR)) (61).

La selección de los estudios que se incluyen en una revisión sistemática se debe hacer por duplicado. Es decir, dos revisores, de forma independiente, revisarán cada una de las referencias identificadas en la búsqueda para decidir si debe ser incluida en la revisión. El proceso de selección se suele hacer en dos fases: en la primera, se revisan los títulos y resúmenes (en inglés: *abstracts*) de las referencias identificadas, esta fase permite excluir sin necesidad de una lectura profunda las referencias que claramente no serán incluidas en la revisión; en la segunda, los estudios que no han sido excluidos en la fase de cribado de título y resumen son revisados en detalle a texto completo. Las discrepancias de criterio entre los revisores se resuelven por consenso o por mediación de un tercer revisor. Los resultados del proceso son reportados mediante un gráfico de flujo (en inglés: *flowchart*) (62,63).

1.3.7. EXTRACCIÓN DE DATOS

La fase de extracción de datos es necesaria para obtener la información relevante de los estudios primarios incluidos en una revisión. La herramienta CHARMS (*CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies*) es una lista con los ítems que deben extraerse de los estudios individuales que se incluyen en una revisión sistemática de estudios de desarrollo, validación o actualización de modelos pronósticos (64). Sin embargo, la pobre e incompleta comunicación de resultados, tanto de los aspectos metodológicos del desarrollo del modelo, tales como el tamaño de la muestra, el número de eventos o el número y manejo de los candidatos predictores, así como de los resultados obtenidos, tales como la ecuación de regresión (incluyendo la constante) o el rendimiento predictivo del modelo, es muy frecuente y dificulta sobre manera la evaluación crítica de los estudios incluidos en una revisión.

1.3.8. EVALUACIÓN DEL RIESGO DE SESGO

La evaluación crítica de la calidad de los estudios identificados es otro aspecto clave de las revisiones sistemáticas. Para los estudios de modelos pronósticos, la herramienta PROBAST (*PRediction model Risk Of Bias ASsessment Tool*) proporciona una lista de ítems señalados para evaluar el riesgo de sesgo en cuatro dominios: participantes, predictores, desenlaces y análisis (65). Muchos modelos pronósticos son clasificados con alto riesgo de sesgo por deficiencias en el diseño y los métodos de análisis estadístico (8,38).

1.3.9. META-ANÁLISIS

El análisis estadístico que combina los resultados de los estudios individuales incluidos en una revisión para producir un resultado global se conoce como meta-análisis. El meta-análisis tendrá sentido solo cuando los estudios identificados sean considerados suficientemente robustos y homogéneos. En estudios de modelos pronósticos, el metaanálisis puede no ser apropiado si los diseños de los estudios son muy diferentes entre sí, si las definiciones del desenlace y/o predictores no son suficientemente homogéneas o si existen dudas acerca de la calidad de los estudios. En investigación sobre factores pronóstico, este resultado global suele ser el resumen de la estimación de una medida de asociación (efecto) (66). En estudios de modelos pronóstico, este resumen suele ser una medida del rendimiento predictivo, en estudios (67). En el estudio incluido en esta tesis, la síntesis fue la combinación de los modelos identificados en un único modelo agregado o meta-modelo (68,69).

En las revisiones sistemáticas de modelos pronósticos el objetivo más común es identificar los modelos disponibles para una misma población y desenlace de interés y resumir cualitativamente sus rendimientos predictivos. Sin embargo, si los modelos son suficientemente homogéneos tanto en el diseño como en las definiciones de los desenlaces y predictores, una alternativa es sintetizar los modelos identificados en un único meta-modelo. Para poder agregar los modelos siguiendo la metodología propuesta por Debray *et al.* se precisa disponer de una muestra de validación con datos primarios (68,69). La síntesis de modelos pronósticos presenta múltiples retos, y recientemente, se ha publicado una guía de para revisiones sistemáticas y meta-análisis de modelos pronóstico (67).

En la sección de resultados, en el apartado de justificación y en la discusión de los aspectos metodológicos del estudio 2 de esta tesis se presentan con más detalle los métodos de agregación de modelos.

1.3.10. COMUNICACIÓN DE LOS RESULTADOS

Como en todos los estudios de investigación una comunicación clara y completa es esencial. En las revisiones sistemáticas las recomendaciones para guiar la adecuada comunicación se basan en PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) (62,63).

1.4. ESCENARIOS CLÍNICOS

Esta tesis de pronóstico se basa en los resultados de un modelo predictivo desarrollado con datos primarios y una revisión sistemática que permitió la creación de un meta-modelo a partir de los modelos existentes. Para aplicar los desarrollos metodológicos llevados a cabo se han seleccionado dos escenarios clínicos. A continuación se describen los mencionados escenarios clínicos sobre los que se quiere realizar predicciones.

1.4.1. ESCENARIO CLÍNICO 1

Las mujeres con epilepsia tienen 10 veces más probabilidades de morir durante el embarazo que aquellas sin la afección (70), y las convulsiones son una causa común de muerte (71). La falta de reconocimiento de las mujeres en situación de alto riesgo por parte de los profesionales de la atención primaria y especializadas se ha destacado como el principal factor detrás de las muertes maternas relacionadas con la epilepsia (71–73). Además, hasta el 40% de las mujeres interrumpen su medicación antiepiléptica durante el embarazo debido a preocupaciones sobre los efectos de los fármacos en el feto, lo que aumenta el riesgo de convulsiones (74,75).

Podrían evitarse muchas muertes maternas en mujeres con epilepsia con la participación oportuna de especialistas en el nivel asistencial adecuado (72). Las crisis epilépticas durante el embarazo también tienen un impacto negativo en la vida diaria. Por ejemplo, el embargo del permiso de conducir durante este periodo podría afectar al empleo, las relaciones sociales y la calidad de vida de estas mujeres (76–78). Las gestantes con epilepsia en riesgo de convulsiones necesitan un plan de seguimiento clínico personalizado para la atención prenatal, intraparto y posnatal, que requiere un abordaje multidisciplinar. Además, las mujeres con alto riesgo de convulsiones necesitan una estrecha monitorización durante el parto con medidas adecuadas de alivio del dolor (79). Sin embargo, la falta de orientación sobre lo que constituye un embarazo de alto riesgo es un factor que ha contribuido a las variaciones en la atención de las mujeres embarazadas con epilepsia (79).

Dadas estas premisas anteriores, se hace necesario una predicción personalizada del riesgo de crisis epilépticas de las mujeres embarazadas en tratamiento antiepiléptico.

1.4.2. ESCENARIO CLÍNICO 2

La endocarditis infecciosa es una enfermedad rara pero asociada con una alta morbilidad y mortalidad y cuyo manejo es a menudo complejo (80,81). El tratamiento de elección para un

porcentaje elevado, entre 20% y 50%, de los pacientes con endocarditis infecciosa es una cirugía cardiaca (82,83).

La cirugía cardiaca de la endocarditis infecciosa consiste en la exéresis y resección de los tejidos afectos por la inflamación e infección que permite la reconstrucción cardiaca de las áreas dañadas, contribuye junto con el tratamiento antibiótico a la erradicación del microorganismo causal y reduce la morbilidad y mortalidad asociadas a la enfermedad (84). Sin embargo, la mortalidad asociada a la cirugía cardiaca en endocarditis infecciosa es alta y varía de forma considerable en función de las características del paciente (85). A pesar de que la cirugía puede ser el tratamiento definitivo de la enfermedad este representa un reto para el cirujano cardiaco por su elevada complejidad.

Las indicaciones de cirugía están definidas en las actuales guías de práctica clínica pero, aunque son ampliamente aceptadas, se establecieron a partir de consenso de expertos, con sentido común pero con poca evidencia científica (86). Por ello, actualmente, existe un gran interés en conocer el pronóstico de los pacientes que pueden ser sometidos a una cirugía por endocarditis infecciosa y esto ha derivado en el desarrollo de múltiples modelos predictivos que pretenden ser útiles en el proceso de toma de decisión. Los investigadores, en lugar de utilizar la información de los modelos predictivos ya desarrollados, y actualizar o modificar estos en caso de ser preciso, desarrollan nuevos modelos partiendo desde cero usando los datos disponibles.

Este escenario en el que coexisten muchos modelos para estimar el riesgo de mortalidad en pacientes con endocarditis infecciosa genera dudas e incertidumbre entre los especialistas. Existe una necesidad no cubierta de disponer de un modelo pronóstico fiable basado en la mejor evidencia disponible que estime predicciones válidas del riesgo de mortalidad postoperatoria en pacientes con endocarditis infecciosa.

Ambas necesidades descritas en esta sección han sido abordadas en los estudios que se presentan en las siguientes páginas.

HIPÓTESIS

2. *HIPÓTESIS*

En un escenario de investigación con suficientes datos primarios,

- el empleo de métodos estadísticos avanzados que combinan procedimientos de imputación múltiple de datos faltantes con técnicas iterativas de selección de predictores y estimación penalizada de sus coeficientes, permite conocer el riesgo de crisis epilépticas durante el embarazo en mujeres con en tratamiento antiepiléptico.

En un escenario de investigación en el que coexisten múltiples modelos pronósticos desarrollados para una misma población diana y para predecir un mismo desenlace de interés,

- la aplicación de métodos sistemáticos de revisión y evaluación del riesgo de sesgo, combinado con técnicas de agregación de los modelos existentes, permite desarrollar un meta-modelo para predecir de forma más precisa el riesgo de mortalidad postoperatoria en pacientes con diagnóstico de endocarditis infecciosa.

En un escenario de investigación en el cual los métodos estadísticos son complejos y requieren de conocimientos estadísticos y de computación avanzados para los usuarios,

- la disponibilidad de una herramienta para realizar la validación interna de un modelo de regresión logística empleando técnicas de remuestreo bootstrapping, facilita a los investigadores el desarrollo de modelos pronósticos más robustos y fiables, así como una comunicación de los resultados del modelo más detallada y estandarizada.

OBJETIVOS

3. OBJETIVOS

OBJETIVO 1.

- Estimar el riesgo de crisis epilépticas durante el embarazo y hasta seis semanas después del parto en mujeres con epilepsia en tratamiento.
- Validar externamente el rendimiento predictivo del modelo en una cohorte independiente de mujeres con un seguimiento de la enfermedad epiléptica heterogéneo.
- Determinar la utilidad del modelo pronóstico en diferentes escenarios de aplicabilidad.
- Crear una calculadora on-line de libre acceso basada en el modelo pronóstico estimado.

OBJETIVO 2.

- Identificar y evaluar de forma crítica la calidad de los modelos predictivos del riesgo de mortalidad postoperatoria en pacientes con diagnóstico de endocarditis infecciosa.
- Sintetizar los modelos pronósticos existentes en un meta-modelo empleando técnicas de agregación y una muestra de validación de un registro nacional de endocarditis.
- Crear una calculadora on-line de libre acceso basada en el meta-modelo pronóstico estimado.

OBJETIVO 3.

- Desarrollar una nueva herramienta de análisis estadístico para el software Stata que permita a los investigadores realizar de forma óptima la validación interna mediante técnicas bootstrapping de modelos pronósticos de regresión logística.

RESULTADOS

ESTUDIO 1

PREDICTING SEIZURES IN PREGNANT WOMEN WITH EPILEPSY: DEVELOPMENT AND EXTERNAL VALIDATION OF A PROGNOSTIC MODEL

4. RESULTADOS

4.1. ESTUDIO 1: PREDICTING SEIZURES IN PREGNANT WOMEN WITH EPILEPSY: DEVELOPMENT AND EXTERNAL VALIDATION OF A PROGNOSTIC MODEL

4.1.1. RESUMEN

Las convulsiones epilépticas son la principal causa de mortalidad en mujeres con epilepsia durante el embarazo. A pesar de este hecho, no se dispone de una herramienta que permita predecir el riesgo de crisis epiléptica en mujeres durante el periodo de gestación.

En este estudio se ha desarrollado y validado externamente un modelo pronóstico para predecir el riesgo de convulsiones epilépticas durante el embarazo y hasta seis semanas después del parto en mujeres en tratamiento antiepiléptico. El modelo emplea predictores cuya información se recoge en la primera visita prenatal. Para promover el uso de este modelo predictivo en la práctica se ha implementado una herramienta on-line para facilitar el cálculo del riesgo de crisis epiléptica dadas las características de la gestante.

Para este estudio se emplearon los datos de 527 mujeres embarazadas con epilepsia en tratamiento incluidas en un estudio de cohortes prospectivo (estudio EMPiRE: *AntiEpileptic drug Monitoring in PREgnancy*) que reclutó mujeres en 50 hospitales del Reino Unido entre noviembre de 2011 y agosto de 2014 (87). Este estudio estaba anidado en un ensayo clínico doble-ciego que comparaba la eficacia y seguridad de dos estrategias de monitorización de los niveles de fármacos antiepilépticos en mujeres embarazadas. El modelo fue desarrollado en una cohorte (n = 399) que incorporó mujeres con un seguimiento clínico de la enfermedad sin monitorización rutinaria de los niveles de fármaco en sangre, y fue validado en una cohorte (n = 128) que incorporó a mujeres que recibieron una asistencia con una monitorización rutinaria de los niveles de fármaco en sangre. El desenlace de interés fue la crisis epiléptica (convulsiones). Se ajustó un modelo usando el método de regresión LASSO (*Least Absolute Shrinkage and Selection Operator*). El rendimiento del modelo fue evaluado en términos de discriminación mediante el estadístico *C*, y en términos de su calibración mediante la pendiente de calibración y la *calibration-in-the-large*. Se determinó el beneficio neto de usar el modelo para varios umbrales de probabilidad, con el objetivo de ayudar en la toma de decisiones clínicas. En la atención al embarazo de mujeres epilépticas, se plantean distintos escenarios de decisión como por ejemplo derivar o no a la gestante a un nivel asistencial especializado, determinar la

frecuencia e intensidad de la monitorización del embarazo, o tomar decisiones relacionadas con los cambios en la medicación antiepiléptica.

De las 399 mujeres de la cohorte de desarrollo y las 128 de la cohorte de validación, 183 (46%) y 57 (45%) sufrieron convulsiones epilépticas, respectivamente. El modelo incluyó la edad en la primera crisis epiléptica, el tipo de crisis epiléptica, antecedentes de enfermedad mental, presencia de crisis tónico-clónicas y no tónico-clónicas en los 3 meses previos al embarazo, ingreso hospitalario por convulsiones en previos embarazos, dosis de lamotrigina y dosis de levetiracetam. El modelo mostró buen rendimiento predictivo para el riesgo de convulsiones epilépticas en mujeres embarazadas en tratamiento antiepiléptico. El estadístico *C* corregido por el exceso de optimismo en la validación interna fue 0.79 (IC 95% 0.75; 0.84). En la validación externa, el modelo mostró un buen rendimiento (estadístico *C* 0.76 (IC 95% 0.66; 0.85) y pendiente de calibración 0.93 (IC 95% 0.44; 1.41) pero con estimaciones imprecisas. En umbrales de riesgo entre 12% y 99%, usar las probabilidades estimadas por el modelo EMPiRE para decidir una acción de atención sanitaria en las gestantes mostró tener un beneficio neto positivo comparado con las estrategias de actuar sobre todas las embarazadas o no hacerlo sobre ninguna. Entre las limitaciones de este estudio se tiene la gran variabilidad de la edad gestacional de las mujeres al reclutamiento, la recogida con carácter retrospectivo de los antecedentes de epilepsia, los potenciales errores de clasificación del tipo de crisis, el tamaño relativamente pequeño de la cohorte de validación. Además, la utilidad clínica del modelo estaba restringida a umbrales de toma de decisión superiores al 12%. Por último, los resultados del modelo no pueden ser generalizados a países de bajos y medios ingresos. A pesar de estas limitaciones, los resultados del estudio sugieren que la integración del uso de esta herramienta predictiva dentro de los procedimientos de la primera visita de la atención al embarazo podría ayudar a personalizar los cuidados en mujeres con epilepsia.

4.1.2. JUSTIFICACIÓN Y ASPECTOS METODOLÓGICOS

Como se ha destacado en la introducción de la tesis, la calidad de los estudios de desarrollo y validación de modelos pronósticos es, en general, pobre. Por lo que es interesante hacer mención específica a algunos de los aspectos metodológicos empleados en el desarrollo del modelo pronóstico EMPiRE.

El estudio EMPiRE incorporó 560 mujeres desde 50 centros de maternidad de Reino Unido. En función de la atención clínica de la epilepsia en las gestantes la muestra fue dividida en una cohorte de desarrollo y otra de validación, por lo que se evitó el uso de técnicas de aleatorización

para la división de la muestra, práctica desaconsejada por su ineficiencia (40). La cohorte de desarrollo incorporó gestantes que no se sometían a monitorización rutinaria de los niveles de fármacos antiepilépticos en sangre, sino que se realizaba una monitorización clínica. Esta estrategia está alineada con los estándares de seguimiento de la epilepsia en gestantes en el Reino Unido. Por su parte, la cohorte de validación incluyó gestantes sometidas a una monitorización rutinaria de niveles de fármaco en sangre, siguiendo los estándares de otros países como Estados Unidos. Por tanto, la separación (*data-split*) de la muestra en los conjuntos de derivación y validación puede ser considerado como una validación externa pues el criterio de división no fue aleatorio sino basado en la asistencia clínica de estas mujeres. De esta manera pudimos determinar la transportabilidad del modelo para distintos contextos de la práctica clínica.

Otro aspecto metodológico interesante que merece mención fue la elección de los candidatos predictores. La elección de los predictores no debe estar basada en la significación estadística de las asociaciones univariables, pues podríamos incluir en el modelo predictores con asociaciones espurias o, al contrario, dejar fuera predictores cuya asociación con el desenlace ha sido probada es estudios previos (29,48). Para el desarrollo del modelo EMPiRE, un equipo multidisciplinar formado por neurólogos, obstetras y metodólogos consensuó los 19 candidatos predictores desde una lista de 65 variables basales disponibles, basados en la evidencia científica y la relevancia en los cuidados clínicos, independientemente de la asociación estadística en análisis univariables.

En cuanto a los análisis estadísticos para el desarrollo del modelo, cabe destacar la complejidad que conlleva la validación interna del modelo empleando técnicas bootstrapping en presencia de métodos de imputación múltiple y de métodos de penalización y selección de predictores (LASSO).

Con el objetivo de poder utilizar la información de mujeres que presentaban datos faltantes en alguno de los predictores en estudio, se emplearon técnicas de imputación múltiple para inferir los datos ausentes utilizando el método de ecuaciones encadenadas (MICE) (88,89). Este proceso consiste en generar múltiples conjuntos de datos imputando los datos faltantes mediante técnicas de regresión.

Para la estimación del modelo se construyeron 10 conjuntos de datos con imputación de los valores faltantes usando MICE (paso 1) (89). Para cada conjunto de imputación se desarrolló un modelo de regresión logística multivariable usando el método LASSO, el cual simultáneamente

selecciona los predictores y los penaliza reduciendo el riesgo de sobreajuste (paso 2) (45). Hicimos un promedio de los coeficientes del modelo usando las reglas de Rubin para obtener los coeficientes finales del modelo (paso 3) (88). Para obtener los intervalos de confianza no paramétricos del 95% para los coeficientes del modelo, repetimos los pasos 2 y 3 anteriores en 1000 muestras bootstrap (100 de cada conjunto de imputación). Los límites del intervalo de confianza del 95% para cada coeficiente fueron los percentiles 2.5 y 97.5 de su distribución (42).

Una vez construido el modelo pronóstico, la validación interna se llevó a cabo de nuevo mediante el empleo de técnicas bootstrapping. En la figura 4 se presenta el esquema de análisis completo seguido en el estudio.

La evaluación del impacto clínico del modelo fue realizada mediante técnicas de análisis de curva de decisión (52). Una regla de decisión debe tomar en consideración las consecuencias (beneficios y daños) de identificar correcta o erróneamente a mujeres que sufrirán una crisis epiléptica. Por ejemplo, si superar cierto umbral de riesgo de convulsiones epilépticas implicara la retirada de la licencia de conducir, esta decisión puede evitar un accidente pero también podría suponer la pérdida del empleo de la mujer. El análisis de curva de decisión permite situar en una balanza las consecuencias de no identificar una mujer de alto riesgo y, por tanto, perder la oportunidad de mejorar la atención clínica de su enfermedad durante el embarazo, y las consecuencias de considerar erróneamente a una mujer como de alto riesgo y sobretratar su enfermedad innecesariamente, cuyo efecto también pueden ser perjudicial para la madre y el bebé.

Aplicar el modelo en la práctica clínica puede resultar complejo y, en ocasiones, los clínicos no saben cómo emplear los coeficientes del modelo para realizar las estimaciones del riesgo dadas las características de la gestante. Esto limita la aplicabilidad del modelo. Un modo de hacer más usable el modelo EMPiRE fue desarrollando un nomograma. Un nomograma es un instrumento semi-gráfico de cálculo que permite estimar rápidamente la probabilidad de tener un evento (en nuestro estudio una crisis epiléptica) a partir de los valores individuales de los predictores de un individuo concreto (en nuestro estudio mujer embarazada en tratamiento antiepiléptico) (20).

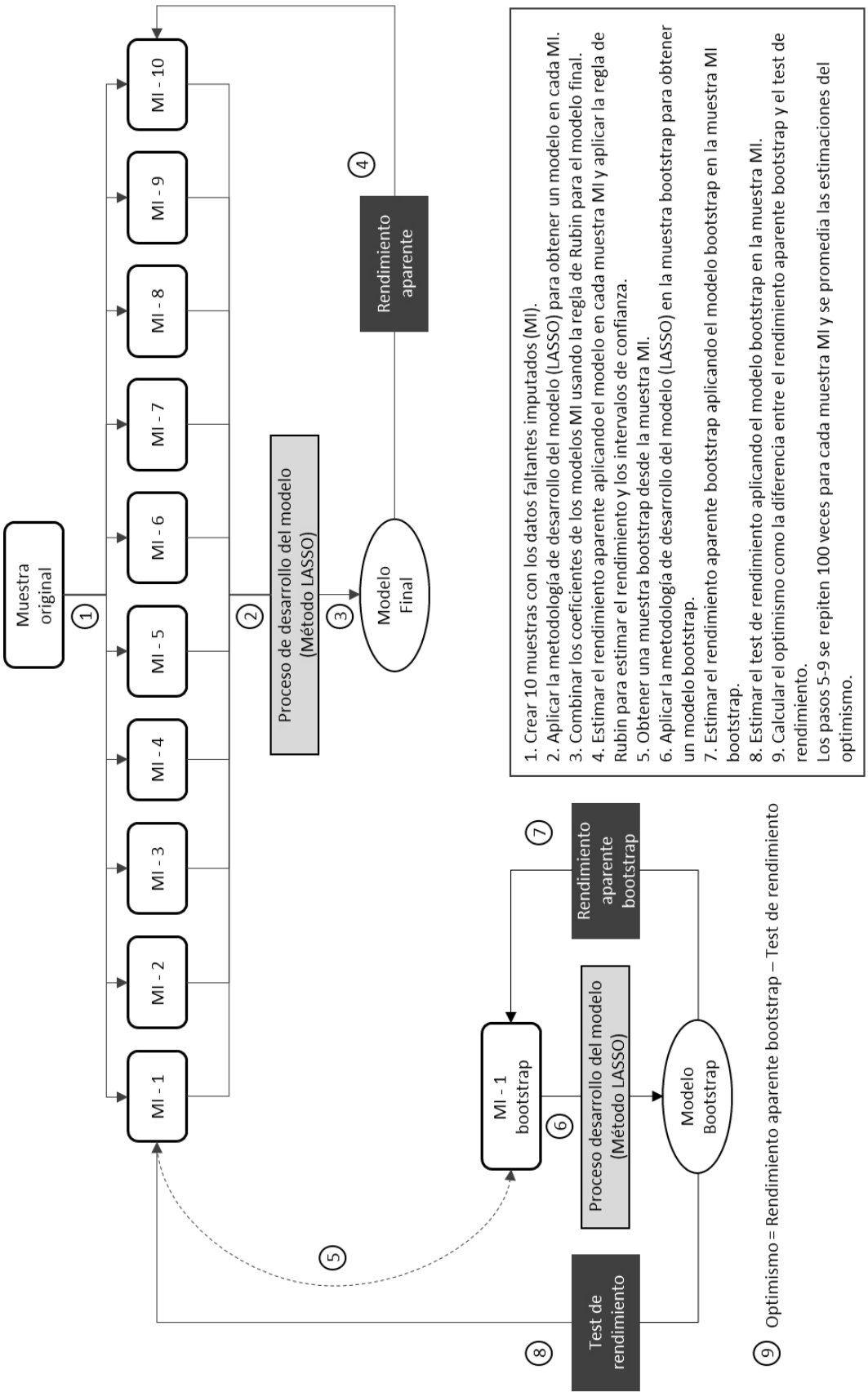


FIGURA 4. ESQUEMA VALIDACIÓN INTERNA BOOTSTRAP EN COMBINACIÓN CON MÚLTIPLE IMPUTACIÓN

Con el fin de facilitar aún más el uso del modelo EMPiRE en la práctica clínica, y la sociedad general, se ha creado una calculadora online que permite, tanto a clínicos como mujeres embarazadas, conocer el riesgo de sufrir convulsiones epilépticas basado en las características de la mujer en la visita prenatal. La calculadora está disponible en la plataforma Evidencio (https://www.evidencio.com/models/show/1799) y en la aplicación móvil del mismo nombre. Evidencio es una plataforma de investigación que proporciona una biblioteca transparente, dinámica y de alta calidad de modelos de predicción clínicamente relevantes.

4.1.3. APLICACIÓN

El empleo de la aplicación es sencillo. Los usuarios, clínicos o mujeres embarazadas, tan sólo tienen que introducir la información referente a la visita prenatal de los predictores incluidos en el modelo. La aplicación reporta la probabilidad de sufrir convulsiones epilépticas junto con una breve interpretación de los resultados.

A continuación, se presenta un ejemplo del uso de la aplicación:

El modelo EMPiRE es aplicado a una mujer embarazada que tuvo su primera crisis epiléptica a los 25 años, las convulsiones epilépticas fueron de tipo no tónico-clónico. El tratamiento antiepiléptico actual (cuando acude a su cita prenatal) consiste en 3000 mg./día de levetiracetam y 400 mg./día de lamotrigina. La mujer ha sufrido alguna crisis epiléptica de tipo no tónico-clónico en los tres meses previos al embarazo, pero ninguna de tipo tónico-clónico. No presenta antecedentes de dificultad en el aprendizaje o desórdenes de salud mental. En embarazos previos la mujer había sido ingresada por crisis epiléptica (Figura 5).

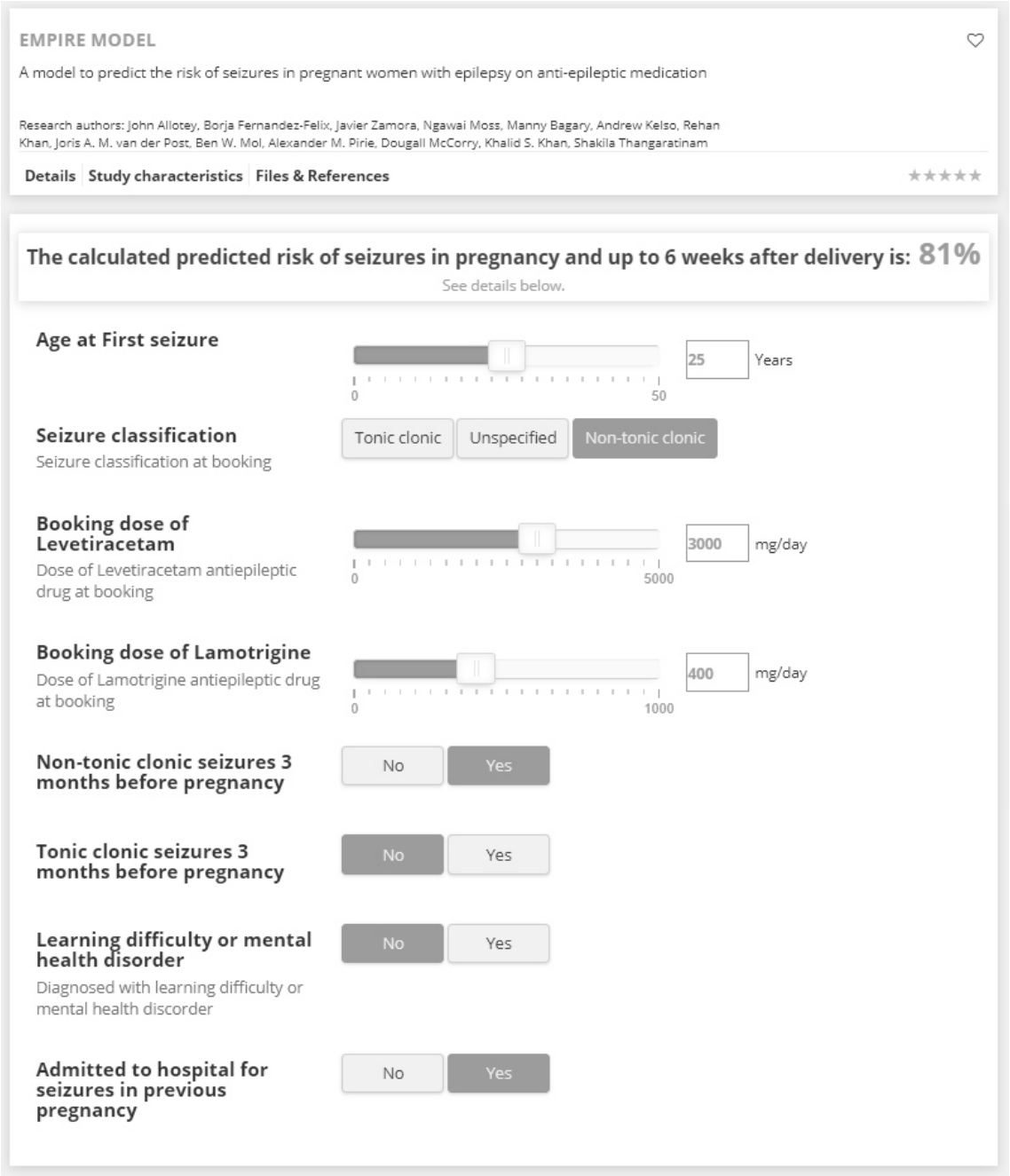


FIGURA 5. EJEMPLO DE APLICACIÓN DEL MODELO EMPIRE EN LA CALCULADORA ONLINE DE EVIDENCIO

La probabilidad de que esta mujer sufra una crisis epiléptica durante su embarazo o en las 6 semanas posteriores al parto es del 81% (Figura 6).

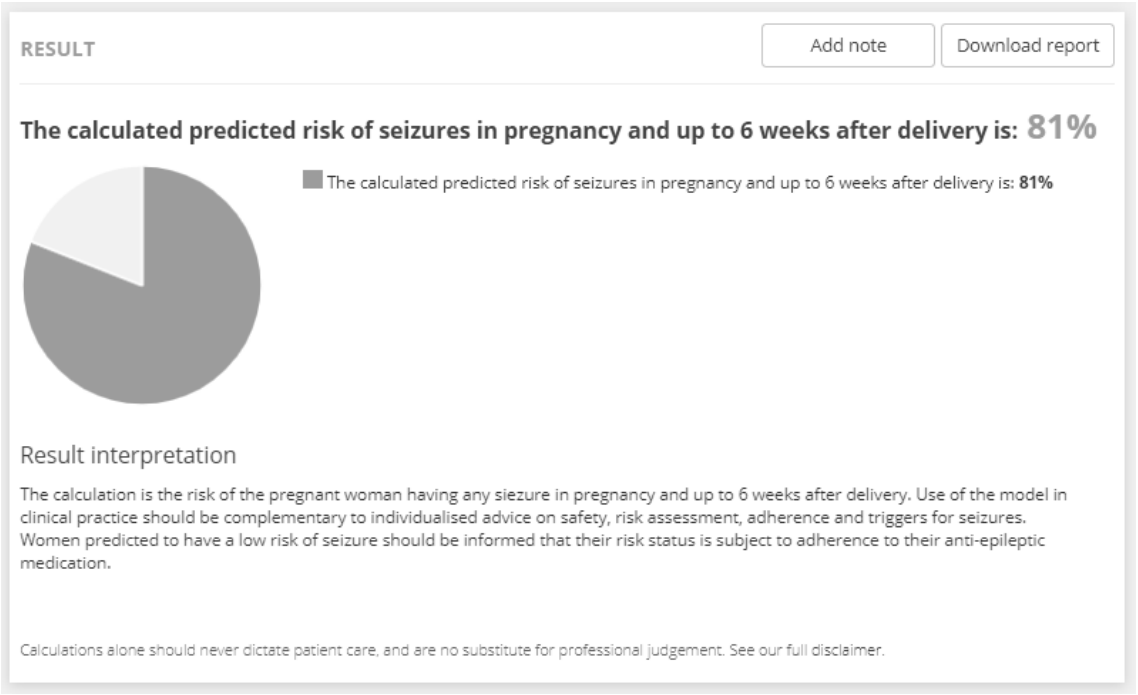


FIGURA 6. RIESGO DE CRISIS EPILÉPTICA ESTIMADO SEGÚN EL MODELO EMPIRE PARA LA MUJER DEL EJEMPLO

4.1.4. ARTÍCULO

Los resultados de este estudio han sido publicados con el título "*Predicting seizures in pregnant women with epilepsy: Development and external validation of a prognostic model*" en la revista PLoS Medicine perteneciente al primer decil de la categoría "Medicine, General and Internal". Enlace: (https://doi.org/10.1371/journal.pmed.1002802).

PLOS MEDICINE

RESEARCH ARTICLE

Predicting seizures in pregnant women with epilepsy: Development and external validation of a prognostic model

John Allotey[1,2], Borja M. Fernandez-Felix[3,4], Javier Zamora[1,3,4]*, Ngawai Moss[5], Manny Bagary[6], Andrew Kelso[7], Rehan Khan[8], Joris A. M. van der Post[9], Ben W. Mol[10], Alexander M. Pirie[11], Dougall McCorry[7], Khalid S. Khan[1,2], Shakila Thangaratinam[1,2]

1 Barts Research Centre for Women's Health, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom, **2** Multidisciplinary Evidence Synthesis Hub, Queen Mary University of London, London, United Kingdom, **3** CIBER Epidemiology and Public Health, Madrid, Spain, **4** Clinical Biostatistics Unit, Hospital Ramón y Cajal, Madrid, Spain, **5** Patient and Public Involvement, Katie's Team, Katherine Twining Network, Queen Mary University of London, London, United Kingdom, **6** Queen Elizabeth Hospital Birmingham, University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom, **7** Department of Neurology, Royal London Hospital, Barts Health NHS Trust, London, United Kingdom, **8** Department of Obstetrics and Gynaecology, Royal London Hospital, Barts Health NHS Trust, London, United Kingdom, **9** Department of Obstetrics and Gynaecology, University of Amsterdam, Academic Medical Centre, Amsterdam, The Netherlands, **10** Monash University, Monash Medical Centre, Clayton, Victoria, Australia, **11** NHS Education for Scotland, Edinburgh, United Kingdom

* javier.zamora@hrc.es

Citation: Allotey J, Fernandez-Felix BM, Zamora J, Moss N, Bagary M, Kelso A, et al. (2019) Predicting seizures in pregnant women with epilepsy: Development and external validation of a prognostic model. PLoS Med 16(5): e1002802. https://doi.org/10.1371/journal.pmed.1002802

Academic Editor: Lucy C Chappell, King's College London, UNITED KINGDOM

Received: November 15, 2018

Accepted: April 11, 2019

Published: May 13, 2019

Data Availability Statement: Data cannot be shared publicly because it belongs to the sponsor Queen Mary University of London and consent was not sought from participants to share publicly. Data are available from the sponsor Institutional Data Access / Ethics Committee (contact via the chairperson, Sally Kerry at pctu-data-sharing@qmul.ac.uk) for researchers who meet the criteria for access to confidential data.

Funding: The EMPiRE trial was funded by a grant from the National Institute of Health Research

Abstract

Background

Seizures are the main cause of maternal death in women with epilepsy, but there are no tools for predicting seizures in pregnancy. We set out to develop and validate a prognostic model, using information collected during the antenatal booking visit, to predict seizure risk at any time in pregnancy and until 6 weeks postpartum in women with epilepsy on antiepileptic drugs.

Methods and findings

We used datasets of a prospective cohort study (EMPiRE) of 527 pregnant women with epilepsy on medication recruited from 50 hospitals in the UK (4 November 2011–17 August 2014). The model development cohort comprised 399 women whose antiepileptic drug doses were adjusted based on clinical features only; the validation cohort comprised 128 women whose drug dose adjustments were informed by serum drug levels. The outcome was epileptic (non-eclamptic) seizure captured using diary records. We fitted the model using LASSO (least absolute shrinkage and selection operator) regression, and reported the performance using C-statistic (scale 0–1, values > 0.5 show discrimination) and calibration slope (scale 0–1, values near 1 show accuracy) with 95% confidence intervals (CIs). We determined the net benefit (a weighted sum of true positive and false positive classifications) of using the model, with various probability thresholds, to aid clinicians in making individualised decisions regarding, for example, referral to tertiary care, frequency and intensity

(NIHR) Health Technology Assessment (HTA) Program (09/55/38). No financial support was received for the prognostic model developed and validated in this article. BWM is supported by an NHMRC Practitioner Fellowship (GNT1082548).

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: AMP has been paid to provide medico-legal reports on the standard of care of women with epilepsy in pregnancy and use of anti-epileptic drugs, and has jointly held grants from the NIHR and Epilepsy Action for research on epilepsy in pregnancy. The other authors have declared that no competing interests exist.

Abbreviations: EEG, electroencephalogram; IPD, individual participant data.

of monitoring, and changes in antiepileptic medication. Seizures occurred in 183 women (46%, 183/399) in the model development cohort and in 57 women (45%, 57/128) in the validation cohort. The model included age at first seizure, baseline seizure classification, history of mental health disorder or learning difficulty, occurrence of tonic-clonic and non-tonic-clonic seizures in the 3 months before pregnancy, previous admission to hospital for seizures during pregnancy, and baseline dose of lamotrigine and levetiracetam. The C-statistic was 0.79 (95% CI 0.75, 0.84). On external validation, the model showed good performance (C-statistic 0.76, 95% CI 0.66, 0.85; calibration slope 0.93, 95% CI 0.44, 1.41) but with imprecise estimates. The EMPiRE model showed the highest net proportional benefit for predicted probability thresholds between 12% and 99%. Limitations of this study include the varied gestational ages of women at recruitment, retrospective patient recall of seizure history, potential variations in seizure classification, the small number of events in the validation cohort, and the clinical utility restricted to decision-making thresholds above 12%. The model findings may not be generalisable to low- and middle-income countries, or when information on all predictors is not available.

Conclusions

The EMPiRE model showed good performance in predicting the risk of seizures in pregnant women with epilepsy who are prescribed antiepileptic drugs. Integration of the tool within the antenatal booking visit, deployed as a simple nomogram, can help to optimise care in women with epilepsy.

Author summary

Why was this study done?

- Pregnant women with epilepsy are at increased risk of death and complications from seizures; their high-risk status during pregnancy and after childbirth is often not recognised.
- Knowledge of an individual's risk of seizures could help healthcare professionals and pregnant women make decisions regarding management.
- To our knowledge, there are currently no models to predict risk of seizures in pregnant women with epilepsy on medication.

What did the researchers do and find?

- We developed the EMPiRE model to predict the risk of seizures in pregnancy and up to 6 weeks after delivery in women with epilepsy on medication whose drug doses were managed based on clinical findings; we validated the model in a separate group of women whose dose management was based on drug levels in the blood.

- The model discriminated well between those with and without seizures, with good agreement between predicted and observed risks across both low- and high-risk women.
- The model is clinically useful for decision-making where the threshold of choice for seizure risk is between 12% and 99%.
- The model showed promising transportability to the validation cohort.

What do these findings mean?

- The EMPiRE prediction model can be used by healthcare professionals to identify pregnant women at high risk of seizures and to plan early referral for specialist input; determine the need for close monitoring in pregnancy, labour, and after childbirth; and assess antiepileptic drug management.
- The performance of the model is unlikely to vary with the antiepileptic drug dose management strategy.

Introduction

Women with epilepsy are 10 times more likely to die in pregnancy than those without the condition [1]—seizures are a common cause of death [2]. Despite warnings from consecutive reports of the Confidential Enquiry into Maternal Deaths (UK) on the failings in antenatal, intrapartum, and postnatal management of women with epilepsy, care of these women remains fragmented [3,4]. A lack of recognition of the women's high-risk status by professionals in primary and in secondary care has been highlighted consistently as the main factor behind epilepsy-related maternal deaths [2,3,5]. Furthermore, up to 4 in 10 women discontinue their antiepileptic medication in pregnancy due to concerns about the effects of drugs on the fetus, thereby increasing their risk of seizures [6,7]. Many maternal deaths in women with epilepsy could be averted with timely specialist input [5]. Seizures in pregnancy also have a negative impact on daily living. For example, the loss of driving license following seizures affects employment, relationships, and quality of life [8–10].

Pregnant women with epilepsy at risk of seizures need a personalised management plan for antenatal, intrapartum, and postnatal care, which requires multidisciplinary input through joint obstetric neurology clinics; however, these clinics are not available in all healthcare centres [11]. Furthermore, women at high risk of seizures need close monitoring in labour, with adequate pain relief measures such as epidural analgesia, and use of long-acting benzodiazepines such as clobazam [11]. Current guidelines recommend the use of these measures in high-risk women [11]. But a lack of guidance on what constitutes high-risk pregnancy is one factor that has contributed to variations in the care of pregnant women with epilepsy [11].

Prediction of seizures based on a woman's individual characteristics not only provides an accurate picture of the risks to inform decision-making, but also promotes effective communication between the multi-specialty teams caring for women with epilepsy. A tool for predicting seizure risk can empower women to make informed decisions on their antenatal and intrapartum care. Furthermore, awareness of one's risk status may lower any anxiety arising from the

unpredictable nature of seizures [12], and promote adherence to medication through risk-informed counselling [13].

To our knowledge, there are currently no models to predict seizure risk in pregnant women with epilepsy. Existing, small retrospective studies provide imprecise estimates of the performance of individual predictors, such as type of seizures and seizure status in pre-pregnancy [14–16]. We aimed to develop and externally validate a prognostic model to predict the risk of seizures in pregnant women with epilepsy on medication, until 6 weeks postpartum. We also planned to determine the net benefit of using the model at various threshold probabilities using decision curve analysis.

Methods

We developed and validated the prognostic model for seizures in the prospective multicentre EMPiRE (AntiEpileptic drug Monitoring in PREgnancy) study, which recruited pregnant women with epilepsy on antiepileptic drugs at first antenatal visit from 50 maternity units in the UK between 4 November 2011 and 17 August 2014 [17]. The UK National Research Ethics Committee approved the EMPiRE study (11/WM/0164), written consent was obtained from participants, and the protocol can be accessed at https://www.journalslibrary.nihr.ac.uk/programmes/hta/095538#/. The research reported here did not require further review by an ethics committee. We reported our prognostic study in line with the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) recommendations, and present our findings as a nomogram, a graphical representation of the model to calculate an individual's risk of seizure [18–20] (S1 TRIPOD Checklist).

Model development and validation cohorts

EMPiRE was a prospective study, and recruited pregnant women with epilepsy on lamotrigine, carbamazepine, phenytoin, or levetiracetam before 24 weeks' gestation. Serum antiepileptic drug levels were assessed every month, but the women and clinicians were blinded to these levels. For women for whom drug levels remained stable, the blinding was maintained (non-randomised cohort) until delivery, and drug doses were adjusted based on clinical features, in line with national recommendations [17,21]. Women whose serum drug levels fell were randomly allocated either to a strategy of adjusting antiepileptic doses based on serum drug levels (after unblinding) or to a strategy of changing the drug doses based on only clinical features (blinding maintained). All participants were followed up until 6 weeks after delivery (S1 Appendix). The study is described in detail elsewhere [17]. For model development, we used the cohort of women who were managed without routine serum drug level monitoring (non-randomised and randomised women), which is in line with standard epilepsy care in the UK [17,21]. We validated the model in the separate cohort of women managed differently, with routine therapeutic drug monitoring, as practised in some countries such as the US, to determine if the model was transportable across varied healthcare practices [22].

Candidate predictors

A multidisciplinary team of neurologists, obstetricians, and researchers selected the candidate predictors for further evaluation in the prognostic model, based on existing evidence and their relevance to clinical care [14,16,23–26]. From an initial list of 65 baseline variables, we selected the following candidate predictors: age at first seizure, history of learning difficulty or mental health disorder, baseline seizure classification (tonic-clonic, non-tonic-clonic, unspecified), history of seizure in the 3 months before pregnancy (tonic-clonic, non-tonic-clonic), number of seizures between start of pregnancy and baseline visit, type of antiepileptic drug taken at

baseline, dose of antiepileptic drug taken at baseline, gestational age at baseline, and hospital admission for seizures in a previous pregnancy. All continuous predictors were assumed to be linearly associated with the outcome.

Outcome

Our main outcome was the occurrence of tonic-clonic (convulsive) or non-tonic-clonic (non-convulsive) seizure [27]. Participants prospectively recorded their epileptic seizures, if any, in purpose-built seizure diaries. To avoid overfitting of multivariable models, the rule of thumb is to ensure that there are 10 events for each predictor variable that was considered for inclusion in the model [28]; we worked within this rule limitation.

We assessed the predictive performance of the model using measures of discrimination (C-statistic) and accuracy (calibration slope). The C-statistic represents the ability of the model to discriminate between those who do and do not experience seizures; a value of 1 indicates perfect discrimination, and a value of 0.5 indicates no discrimination beyond chance [29]. Models are considered to have a good performance when the C-statistic exceeds 0.7 [30]. Calibration refers to agreement between the predicted and observed risk of seizure for all groups of predicted probabilities. A well-calibrated model will have a calibration slope of 1, and all groups will fit close to this line.

Statistical analysis

Model development. Missing values were imputed using 10-fold multiple imputation by chained equations. Within each imputed dataset we developed a multivariable logistic regression model using the LASSO (least absolute shrinkage and selection operator) method, which simultaneously selects the variables and penalises the model coefficients for over-optimism. We selected the lambda parameter that minimised expected model deviance. Final coefficients were combined across imputed datasets using Rubin's rule [31,32]. Confidence intervals for model coefficients were obtained by bootstrap sampling. Bootstrap validation was carried out to adjust the performance of the model for optimism. We repeated the entire modelling process on 100 bootstrap samples drawn from each of the 10 imputed datasets. We did not consider the number of tonic-clonic seizures between the diagnosis of pregnancy and baseline visit as a variable, as it was highly correlated with the predictor variable of tonic-clonic seizure history in the 3 months before pregnancy.

We assessed the overall discriminatory ability of the model using the C-statistic (summarised as the area under receiver operating characteristic curve [AUC ROC]) with 95% confidence interval (CI). Model calibration was visually assessed with a calibration plot representing deciles of predicted probability of seizure against the observed rate in each risk group. The calibration slope was also calculated, which is the slope of the regression line fitted between predicted and observed risk probabilities on the logit scale, with 1 being the ideal value.

External validation. We externally validated the final model in a separate cohort of women whose drug levels were routinely monitored (therapeutic drug monitoring) to plan drug dose adjustments. We report the predictive performance in this cohort using the same measures of discrimination and calibration as used in the model development cohort [33]. We visually assessed the model's calibration by plotting quartiles of predicted probability of seizure against the observed rate in each quartile, and estimated the calibration slope [34,35].

Sensitivity analysis. We undertook sensitivity analysis by combining all available data in the development and validation cohorts to determine if there was any change in model

performance. We evaluated antiepileptic drug dose monitoring strategy (therapeutic drug monitoring versus clinical features monitoring) as a predictor while developing the combined model.

Decision curve analysis. We performed decision curve analysis to assess the clinical value of the model. A decision rule should take into consideration the identification of women likely to have seizures because of the significant consequences (maternal death, accidents, loss of driving license), and the avoidance of unnecessary interventions leading to adverse impact on mother or baby. We determined the net benefit of the model across a wide range of threshold probabilities, instead of simply classifying all pregnant women with epilepsy as predicted to have seizures or classifying no women as predicted to have seizures [36]. We represent the net benefit as a function of the decision threshold in a decision curve plot.

The choice of thresholds will vary according to the planned intervention, the preference of the clinician, and the preference of the mother. For example, if the intervention involves referral of a pregnant woman with epilepsy to a tertiary unit for antenatal and intrapartum care, at a low threshold, false negatives are minimised at the expense of unnecessary referrals to tertiary care. At a high threshold, fewer women are referred, but women who are likely to benefit may be denied access to tertiary care. We expect an intermediate range of thresholds to be clinically acceptable. But if the planned intervention is to increase the dose and number of antiepileptic drugs—which can have adverse side effects on the mother and increase the risk of long-term neurodevelopmental problems in the child—clinicians may choose a higher threshold than what was chosen for tertiary referral.

Nomogram. We also developed a simple, easy-to-use nomogram to calculate the predicted probability of seizures in pregnant women at the time of antenatal booking.

The analyses were performed using Stata software version 15.1 and R software version 3.3.2 [37,38].

Results

The EMPiRE study recruited 560 pregnant women. The model development cohort included 399 women; the validation cohort included 128 women (S1 Appendix).

Characteristics of the women

The average gestational age at baseline was 16.6 weeks (SD 4.0) in the development cohort and 14.9 weeks (SD 4.4) in the validation cohort (Table 1). The mean age at first seizure was similar in both cohorts, at 16 years, and 10%–15% of women had a learning difficulty or mental health disorder. A similar proportion of women in the development and validation cohorts were previously diagnosed to have tonic-clonic seizures (development, 39%; validation, 36%). Overall, 46% (182/399) of women in the development cohort and 39% (50/128) in the validation cohort had experienced seizures in the 3 months prior to pregnancy. Lamotrigine was the commonest antiepileptic drug prescribed in both cohorts; more than half of the women took lamotrigine in the development (226/399, 57%) and validation cohorts (80/128, 63%).

Model performance

Overall, 46% (183/399) of women in the development cohort experienced 1 or more seizures at any time from baseline until 6 weeks after delivery. Tonic-clonic seizures accounted for half of all seizures (90/183, 49%). Eight predictors were significantly associated with seizures and were included in the final multivariable model: age at first seizure, history of mental health disorder or learning difficulty, baseline seizure classification (tonic-clonic, non-tonic-clonic, unspecified), hospital admission for seizures in a previous pregnancy, tonic-clonic seizure in

Table 1. Details of women's characteristics in the development and validation cohorts of the EMPiRE prediction model and the proportion with missing data.

Characteristic	Development cohort (n = 399)		Validation cohort (n = 128)	
	Mean (SD) or n (%)	Number with missing data, n (%)	Mean (SD) or n (%)	Number with missing data, n (%)
Gestational age at baseline (weeks)	16.6 (4.0)	0	14.9 (4.4)	0
History of learning difficulty or mental illness	50 (13%)	1 (0.3%)	20 (16%)	0
Age at first seizure (years)	16.5 (7.4)	5 (1.3%)	16.8 (7.8)	0
≤10 years	70 (17.8%)		28 (21.9%)	
11–20 years	215 (54.6%)		59 (46.1%)	
21–30 years	94 (23.9%)		35 (27.3%)	
31–40 years	15 (3.8%)		6 (4.7%)	
Admission to hospital for seizures in previous pregnancy	28 (7.0%)	22 (5.5%)	14 (10.9%)	7 (5.5%)
Seizure classification at baseline				
Tonic-clonic	155 (39%)	0	46 (36%)	0
Non-tonic-clonic	232 (58%)	0	78 (61%)	0
Unspecified	12 (3%)	0	4 (3%)	0
Seizure in the 3 months before pregnancy	182 (46%)		50 (39%)	
Tonic-clonic	52 (13%)	83 (20.8%)	12 (9%)	32 (25.0%)
Non-tonic-clonic	130 (33%)	0	38 (30%)	0
Number of seizures in pregnancy prior to the baseline visit				
Tonic-clonic	0.7 (2.8)	82 (20.6%)	0.4 (2.1)	31 (24.2%)
Non-tonic-clonic	11.6 (108.4)	0	18.1 (83.4)	0
Antiepileptic drug intake at baseline				
Carbamazepine	74 (19%)	0	16 (13%)	0
Lamotrigine	200 (50%)	0	66 (52%)	0
Levetiracetam	99 (25%)	0	31 (24%)	0
Phenytoin	0	0	1 (1%)	0
Lamotrigine and carbamazepine	1 (0.3%)	0	—	0
Lamotrigine and levetiracetam	25 (6%)	0	14 (11%)	0
Baseline dose of antiepileptic drugs (mg/day)				
Carbamazepine	706.0 (348.5)	0	612.5 (346.2)	0
Lamotrigine	272.1 (155.6)	0	269.4 (160.6)	0
Levetiracetam	1,641.3 (886.8)	0	1,533.3 (760.5)	0
Phenytoin	0	0	200 (—)	0

https://doi.org/10.1371/journal.pmed.1002802.t001

the 3 months before pregnancy, non-tonic-clonic seizure in the 3 months before pregnancy, baseline dose of lamotrigine, and baseline dose of levetiracetam (Table 2). The model is presented as a graphical calculator (nomogram) in Fig 1.

The equation of the EMPiRE prediction model for risk of seizures during pregnancy and until 6 weeks after delivery in women with epilepsy on antiepileptic drugs was as follows:

$$\text{probability(seizure)} = \exp(Y)/(1+\exp(Y))$$

where $Y = -1.39 + (-0.02 * \text{age at first seizure}) + 0.61$ [unspecified seizures] + 0.75 [non-tonic-clonic seizures] + (0.02 * dose of levetiracetam/100) + (0.29 * dose of lamotrigine/100) + 0.66 [non-tonic-clonic seizures in the 3 months before pregnancy] + 1.97 [tonic-clonic seizures in the 3 months before pregnancy] + 0.67 [learning difficulty or mental health disorder] + 0.17 [admitted to hospital for seizures during previous pregnancy].

Table 2. Multivariable LASSO logistic regression of seizure risk prediction in pregnant women with epilepsy.

Candidate predictor	Multivariable analysis after MI (n = 399)	
	OR	Bootstrap 95% CI
Age at first seizure (years)	0.98	0.97, 0.99
History of learning difficulty or mental illness	1.96	1.68, 2.89
Seizure classification at baseline (ref. tonic-clonic)		
Non-tonic-clonic	2.11	1.88, 2.62
Unspecified	1.85	1.64, 4.30
Tonic-clonic seizure in the 3 months prior to pregnancy	7.20	6.63, 11.93
Non-tonic-clonic seizure in the 3 months prior to pregnancy	1.94	1.71, 2.38
Baseline dose of lamotrigine (×100 mg/day)	1.34	1.30, 1.44
Baseline dose of levetiracetam (×100 mg/day)	1.02	1.01, 1.03
Admitted to hospital for seizures in previous pregnancy	1.19	1.08, 1.92
Baseline dose of carbamazepine (×100 mg/day)	—	—
Number of non-tonic-clonic seizures since the start of pregnancy	—	—
Gestational age at baseline (weeks)	—	—

95% CI: Bootstrap limits of the confidence interval obtained from percentiles 2.5 and 97.5. Missing values were imputed using 10-fold MI by chained equations (step 1). We fitted a regression model using the LASSO strategy in each of the 10 imputed datasets (step 2). We averaged model coefficients using Rubin's rule to get the final model coefficients (step 3). To obtain non-parametric 95% confidence intervals for model coefficients, we repeated the previous step 2 and step 3 on 1,000 bootstrap samples. Limits of the 95% confidence interval for each coefficient were the 2.5th and 97.5th percentiles of their distribution.

MI, multiple imputation; OR, odds ratio.

https://doi.org/10.1371/journal.pmed.1002802.t002

All variables were coded as binary (1 when present and 0 when absent) except for age at first seizure (years), dose of lamotrigine (mg/day), and dose of levetiracetam (mg/day).

The apparent C-statistic for the model was 0.80 (95% CI 0.76, 0.85). After bootstrap adjustment for optimism, the final prediction model had a C-statistic of 0.79 (95% CI 0.75, 0.84) to discriminate between women with and without seizures (Table 3). The optimism-adjusted calibration plot (Fig 2) showed mostly good agreement between the predicted and observed risks, and the calibration slope was 1.26 (95% CI 0.98, 1.54).

Our sensitivity analysis, which combined all available data (n = 527), resulted in a model with the same predictors and similar coefficients as the EMPiRE model developed using only the development cohort. The antiepileptic drug monitoring strategy was not found to be a significant predictor of seizures and was therefore not selected in the combined model. The C-statistic and calibration slope of the combined model were 0.78 (95% CI 0.74, 0.82) and 1.22 (95% CI 0.44, 1.46), respectively (S2 Appendix).

External validation and predictive performance

In the external validation cohort, 45% (57/128) of women experienced seizures at any time from baseline until 6 weeks after delivery; tonic-clonic seizures were reported in 39% (22/57) of women who had seizures. The final model showed good discrimination when externally validated, with a C-statistic of 0.76 (95% CI 0.66, 0.85). The model showed mostly good agreement between the predicted and observed risks, with a calibration slope of 0.93 (95% CI 0.44, 1.41) (Fig 2).

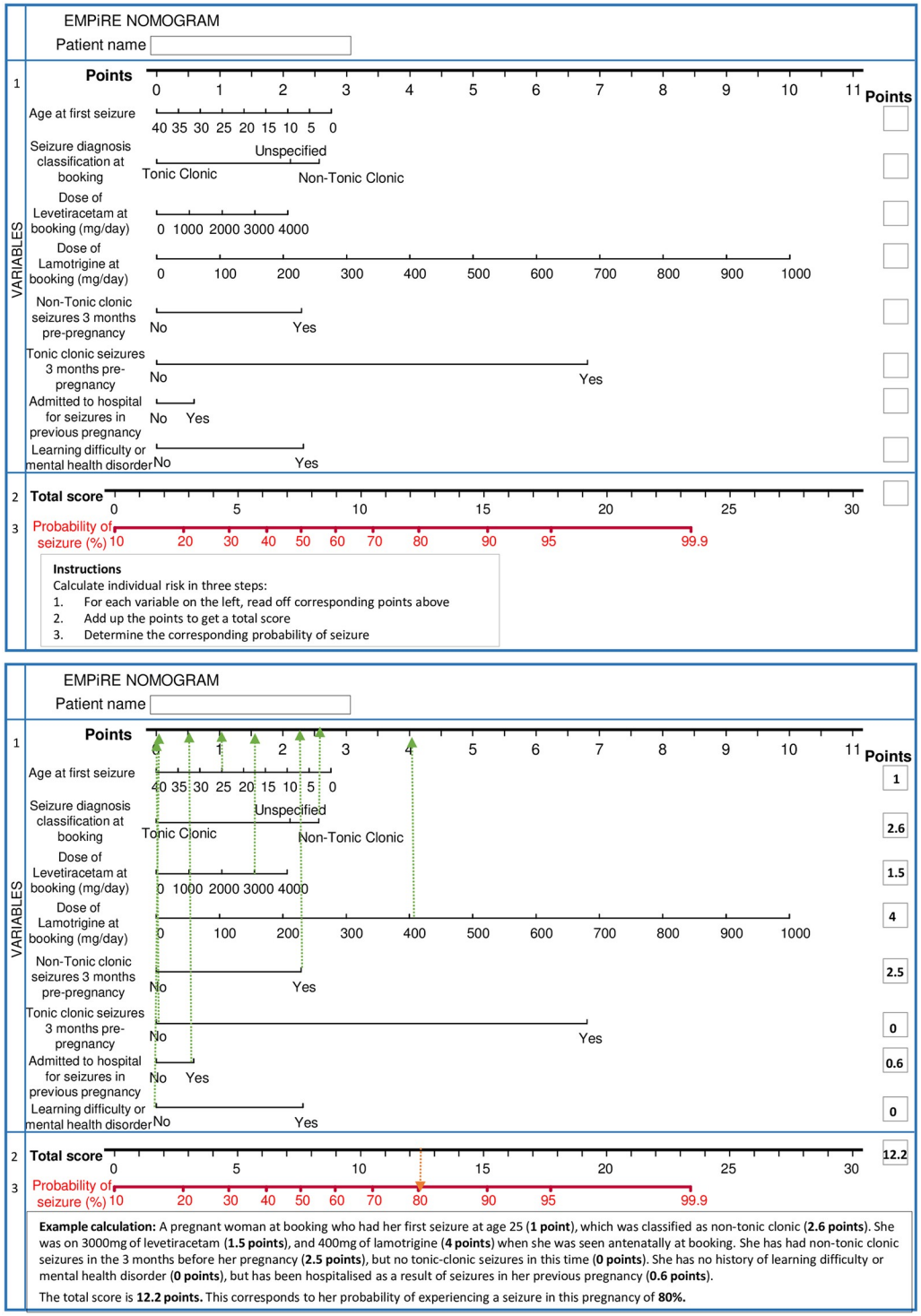


Fig 1. EMPiRE nomogram for predicting the risk of seizures at antenatal booking in pregnant women with epilepsy on antiepileptic drugs: A worked example.

https://doi.org/10.1371/journal.pmed.1002802.g001

Net benefit of model use

In our decision curve analysis (Fig 3), the curve for the EMPiRE model showed positive net benefit for predicted probability thresholds between 12% and 99% compared to managing

Table 3. EMPiRE model performance.

Performance measure	Development cohort (n = 399)	Validation cohort (n = 128)
C-statistic	0.79 (95% CI 0.75, 0.84)	0.76 (95% CI 0.66, 0.85)
Calibration slope	1.26 (95% CI 0.98, 1.54)	0.93 (95% CI 0.44, 1.41)

https://doi.org/10.1371/journal.pmed.1002802.t003

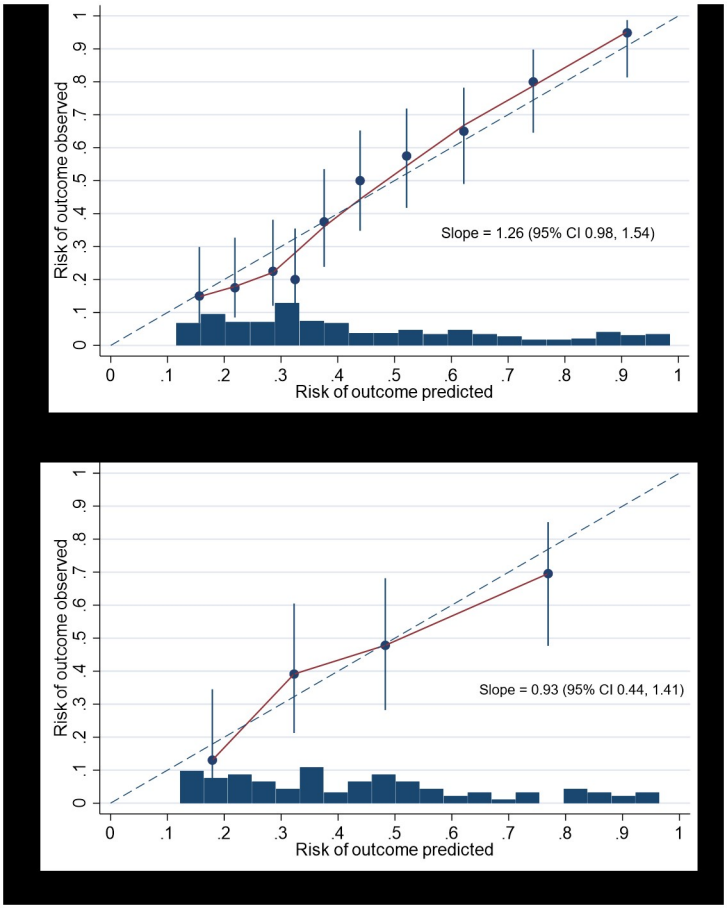


Fig 2. Calibration of the EMPiRE prediction model by comparing observed versus predicted risk of seizures in pregnant women on antiepileptic drugs, with a frequency histogram. Top panel = development cohort; bottom panel = validation cohort.

https://doi.org/10.1371/journal.pmed.1002802.g002

pregnant women with epilepsy as if they will all have seizures or managing them as if none of them will have seizures (i.e., treat-all or treat-none strategies). Table 4 provides estimates of the net benefit of using the model for various probability thresholds. For low thresholds, below 12%, there was no difference between using the EMPiRE model and treating women as if they will all have seizures.

Discussion

Summary of the findings

The EMPiRE model performs well in predicting the risk of seizures at the time of antenatal booking in pregnant women with epilepsy who are prescribed antiepileptic medication. The

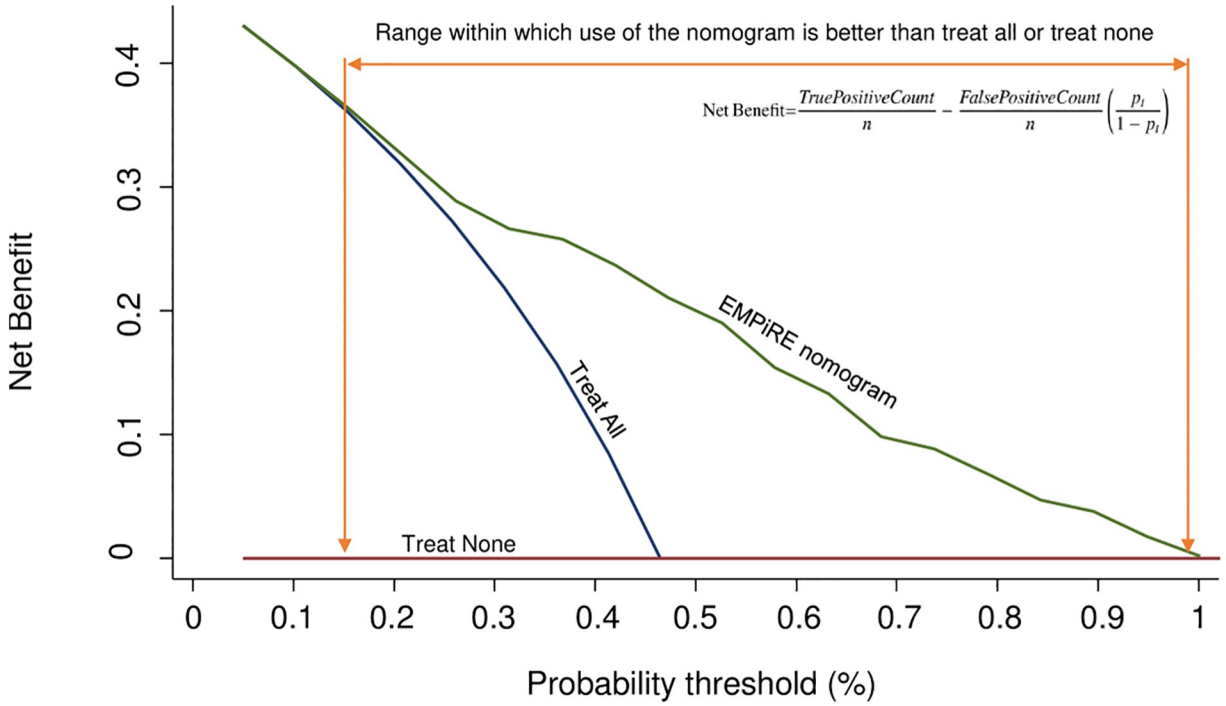


Fig 3. Decision curve analysis using the EMPiRE seizure risk prediction model. Red line (treat none) = net benefit when we assume that no pregnant woman with epilepsy will have the outcome (seizure in pregnancy); blue line (treat all) = net benefit when we assume that all pregnant women with epilepsy will have the outcome; green line (EMPiRE nomogram) = net benefit when we manage pregnant women with epilepsy according to the predicted risk of the outcome (seizure in pregnancy) estimated by the EMPiRE model. The preferred strategy is the one with the highest net benefit at any given threshold.

https://doi.org/10.1371/journal.pmed.1002802.g003

model incorporates routinely available characteristics that are easy to measure, such as age at first seizure, type of seizures, seizures in the 3 months before pregnancy, mental health, admission to hospital for seizures during a previous pregnancy, and dose of antiepileptic drugs. The model is clinically useful over a range of threshold probabilities, and is relevant to general practitioners, epilepsy specialists, obstetricians, and midwives in identifying high-risk women. The model shows potential for transportability across risk groups to settings where routine therapeutic drug monitoring is undertaken, but the findings should be interpreted with caution due to the small number of events in the validation sample. Our simple nomogram is designed to facilitate the model's use in clinical practice.

Strengths and limitations

To our knowledge, ours is the only clinical prognostic model to predict seizures in pregnant women with epilepsy. We developed the model using data from a prospective, high-quality, multicentre study. We evaluated predictors that were clinically relevant and routinely available to healthcare professionals, so that the model can be easily applied in clinical practice. Missing values of predictors were dealt with by multiple imputation, thereby avoiding loss of useful information [39,40]. We developed the model to predict seizures not only in pregnancy, but up to 6 weeks after delivery, a period with increased risks to the mother and baby [11,41]. We adjusted for optimism and addressed issues around overfitting in the model. In addition to providing the model as an easy-to-use nomogram, we provided information on its clinical use at various threshold probabilities for decision-making [35]. The model includes clinical

Table 4. Net benefit of using the EMPiRE prediction model compared to managing women with epilepsy assuming all of them will have seizures in pregnancy or the postpartum period.

Threshold probability	Net benefit		Advantage of using the model	
	Treat all women	EMPiRE model	Difference in net benefit	Reduction in number who do not need the intervention per 100 women
0.05	0.430	0.430	0	0
0.1	0.398	0.398	0	0
0.15	0.363	0.368	0.003	2
0.2	0.323	0.329	0.008	3
0.25	0.278	0.299	0.018	6
0.3	0.227	0.276	0.049	11
0.35	0.167	0.266	0.100	19
0.4	0.098	0.248	0.150	22
0.45	0.016	0.213	0.207	25
0.5	−0.083	0.193	0.286	29
0.55	−0.203	0.166	0.372	30
0.6	−0.353	0.164	0.503	34
0.65	−0.547	0.120	0.663	36
0.7	−0.805	0.111	0.911	39
0.75	−1.165	0.088	1.253	42
0.8	−1.707	0.073	1.774	44
0.85	−2.609	0.057	2.668	47
0.9	−4.414	0.030	4.454	49
0.95	−9.827	0.023	9.852	52
0.99	−53.135	0.000	53.135	54

https://doi.org/10.1371/journal.pmed.1002802.t004

variables that are easily accessible at the time of booking for incorporation into an app or integration into computer systems within healthcare services. A convenient and easy-to-use nomogram of the EMPiRE model allows for immediate use of the model to predict the risk of seizures without the need to remember the formulae behind it.

Our model development and validation approach took into account the significant variations in the management of antiepileptic drug dosages in pregnancy to prevent seizures [42]. While the American Academy of Neurology recommends routine serum therapeutic drug monitoring, with dosage increased if the serum drug level falls [22], the UK National Institute for Health and Care Excellence, Royal College of Obstetricians and Gynaecologists, and Scottish Intercollegiate Guidelines Network guidelines do not recommend routine drug monitoring but drug dose adjustments based mainly on clinical features [11,21,43]. In this paper, we developed and internally validated the model in women whose drug dose was managed based on clinical features to determine the accuracy and reproducibility of the model. Through our external validation, we assessed the transportability of the model to women managed using a different management strategy (therapeutic drug monitoring), which appears promising [44]. Furthermore, when we developed the combined model by using all available data (including women routinely and not routinely monitored for drug levels) in our sensitivity analysis, we did not observe any differences either in the number and type of predictors or in the model's performance compared to the model developed using only women in the development cohort. The antiepileptic drug dose monitoring strategy was not identified to be a significant predictor in the combined model, implying that the model is generalisable irrespective of the strategy.

There are some limitations to this study. The cohorts consisted of women recruited with pre-specified criteria, which may limit the use of the model in all women [19,45]. We did not

include women on sodium valproate; this is consistent with current recommendations against valproate use in pregnancy, due to the increased risk of birth defects and neurodevelopmental disorders [46]. The model can only be used in women managed on phenytoin, lamotrigine, levetiracetam, or carbamazepine and when information is available on all predictors in the context of care in a high-income setting. This limits its transportability to low- and middle-income countries with resource constraints and non-availability of these drugs [47]. Women were recruited at varied gestational ages. However, we evaluated gestational age as a predictor, and it was not selected by the modelling strategy in the final model. The model included history of seizures in the 3 months before pregnancy obtained through retrospective recall, with resultant bias. We consider this to reflect the real life scenario, where women who do not receive pre-pregnancy specialist epilepsy care often do not maintain a prospective seizure diary prior to seeing an epilepsy specialist in pregnancy. It is possible that a different predictor such as history of seizures in the 9 months before pregnancy instead of the 3-month history in our model may have improved its performance [48,49]. We only included clinical predictors routinely available at the time of antenatal booking, and did not evaluate other tests such as electroencephalogram (EEG) or MRI, or risk factors such as history of nocturnal or prolonged seizures, which may be available in specialist epilepsy care. We could not assess any changes in antiepileptic medication before conception because this information was not routinely recorded at antenatal booking, and was not collected in the EMPiRE trial. These additional variables may have improved the performance of the model. Due to the small sample size and the small number of events in the validation cohort (<100), we were limited in our interpretation of the transportability of the model [50].

Comparison to existing evidence

To our knowledge, 2 other prediction models exist for seizures, both involving non-pregnant individuals: seizure prediction in children and adults who have recently stopped their antiepileptic drugs, reported using individual participant data (IPD) meta-analysis, and prediction of subsequent seizures after a single seizure in individuals without clear indication to commence treatment (MESS study) [51,52]. Some predictors in the IPD meta-analysis model such as the age at onset of epilepsy were also present in our model. It is not appropriate or feasible to apply other variables, such as seizure-free interval before antiepileptic drug withdrawal and epileptiform abnormality on EEG, to the pregnant population [51]. The performance of our EMPiRE model was better than that of the IPD meta-analysis model (C-statistic 0.65) [51]. The MESS study, which used split sample validation, did not report the performance of the model with currently recommended measures such as C-statistic and calibration slope, and hence we are unable to compare that model with the EMPiRE model.

Other individual studies such as the EURAP (European and International Registry of Antiepileptic Drugs and Pregnancy) have reported on the association between maternal risk factors and seizures in pregnancy—none provided multivariable prognostic models [14,16]. Compared to a third of pregnant women with epilepsy on medication developing seizures in the EURAP study, 46% of women in the EMPiRE cohorts experienced seizures in pregnancy or until 6 weeks after delivery [14,23]. This difference could be attributable to the known increase in seizures that occurs after delivery in new mothers, as EURAP did not include postnatal mothers or prospective seizure diaries [41]. Inclusion of a selective group of women in the EMPiRE study may also have contributed to the difference. Similarly to the EURAP study, our final model identified lamotrigine dose to be a predictor of seizures [14]. Another small retrospective study identified pre-pregnancy seizure status to be the main predictor of seizures in pregnancy [16], which was also the strongest predictor of seizures in our model.

Relevance to clinical care

Currently pregnant women with epilepsy are managed by varied healthcare professionals such as general practitioners, obstetricians and midwives, and epilepsy specialists. There is no clear pathway for multidisciplinary communication. Joint obstetric neurology clinics are not available in half of the maternity units in UK, a major hindrance for integration of epilepsy care within antenatal care [53]. The first step towards achieving integrated care is effective risk communication of the mother's seizure status. Such a risk-based approach using quantified risk estimates can help to avoid maternal deaths such as those reported in the MBRRACE-UK (Mothers and Babies: Reducing Risk through Audits and Confidential Enquiries across the UK) report [5], where women with epilepsy were never treated by an epilepsy specialist in pregnancy, and were left unmonitored during their hospital admissions without specialist input [5].

We refrained from recommending specific decision thresholds for various interventions as these are likely to vary with the potential adverse effects and costs of the planned intervention. For example, primary care clinicians may consider a 20% cutoff, a level of risk associated with driving restrictions, to be appropriate to make decisions on early referral to tertiary units with joint obstetric neurology clinics. However, secondary care clinicians may choose a higher threshold when the intervention involves frequent antenatal monitoring (weekly or fortnightly), intrapartum use of invasive interventions for pain relief such as epidural and other medications (such as clobazam, which carries a risk of neonatal respiratory depression), or close monitoring in the postnatal period. The choice of threshold in a clinical setting is also likely to vary depending on the epilepsy syndrome and seizure types. For example, the intervention threshold may be lower for patients who tend to experience convulsive seizures than for those who experience absence seizures.

Women's choice of thresholds may depend on the additional time and resources required (for example, long-distance travel to access tertiary care) and the perceived risks to themselves and their babies from the various interventions. If the ability to drive is crucial to the mother for her job and other responsibilities, after discussion with clinicians, she may opt for a lower threshold for interventions in secondary care. But if minimising the risk of long-term adverse offspring neurodevelopmental outcomes is valued more by the mother than minimising the risk of seizures, she may choose a higher threshold for increasing the dose and number of antiepileptic drugs. Our decision curve analysis shows that the model is useful across a wide range of threshold probabilities.

Use of the model in clinical practice should be complementary to individualised advice on safety, risk assessment, drug adherence, and triggers for seizures. Awareness of seizure risk can minimise non-adherence to medication in pregnancy, one of the major factors behind seizure deterioration in pregnancy [6,7,13]. Women predicted to have a low risk of seizure by the model should be informed that their risk status is subject to adherence to their antiepileptic medication. The EMPiRE model does not identify women below 12% risk. Women and clinicians should be aware of this limitation if the probability threshold to make decisions on eligibility for home or water birth is below this threshold.

Relevance to research

The effect of the addition of other markers, such as EEG findings or historical MRI brain imaging reports, on the performance of the model needs further evaluation. There is a need for multiple external validations across different settings and populations to fully appreciate the transportability of the model [54]. The impact of using the EMPiRE model in clinical practice needs to be evaluated through cluster-randomised trials, to assess whether it helps improve

outcomes such as the seizure-free period or quality of life of these women. While the tool is expected to improve women's knowledge of their risk status for seizures in pregnancy, the effect of the EMPiRE model on women's anxiety levels is not known and needs to be assessed. Further studies are needed to assess the acceptability of the tool to women with epilepsy and to healthcare providers, their preferred thresholds of choice, and the cost utilities of consequences of decisions for various false positive and false negative cases.

Conclusions

The EMPiRE nomogram is a simple 8-item prediction tool to calculate the individualised risk of seizures at antenatal booking in pregnant women with epilepsy on antiepileptic drugs. The estimates can help guide individually tailored choices made by patients and clinicians, which may influence the intensity of monitoring in pregnancy and after delivery, place of care, and antiepileptic drug dose adjustment strategy. The model is not clinically useful for decision-making at very low thresholds.

Supporting information

S1 Appendix. Flow diagram of EMPiRE trial participants contributing to model development and external validation.
(TIF)

S2 Appendix. Result of sensitivity analysis combining all available data.
(DOCX)

S1 TRIPOD Checklist.
(DOCX)

Acknowledgments

We are grateful to the original team involved with the collection of the EMPiRE trial data and all women who took part in the study.

Author Contributions

Conceptualization: Ngawai Moss, Khalid S. Khan, Shakila Thangaratinam.

Data curation: John Allotey, Borja M. Fernandez-Felix, Khalid S. Khan, Shakila Thangaratinam.

Formal analysis: John Allotey, Borja M. Fernandez-Felix, Javier Zamora.

Funding acquisition: Khalid S. Khan, Shakila Thangaratinam.

Investigation: John Allotey, Javier Zamora, Manny Bagary, Andrew Kelso, Rehan Khan, Ben W. Mol, Alexander M. Pirie, Dougall McCorry, Khalid S. Khan, Shakila Thangaratinam.

Methodology: John Allotey, Borja M. Fernandez-Felix, Javier Zamora, Joris A. M. van der Post, Ben W. Mol, Dougall McCorry, Khalid S. Khan, Shakila Thangaratinam.

Project administration: John Allotey, Shakila Thangaratinam.

Resources: John Allotey, Khalid S. Khan, Shakila Thangaratinam.

Software: Borja M. Fernandez-Felix.

Supervision: Manny Bagary, Andrew Kelso, Rehan Khan, Ben W. Mol, Alexander M. Pirie, Dougall McCorry, Khalid S. Khan, Shakila Thangaratinam.

Validation: John Allotey, Borja M. Fernandez-Felix, Javier Zamora, Shakila Thangaratinam.

Visualization: John Allotey, Borja M. Fernandez-Felix, Javier Zamora, Ngawai Moss, Khalid S. Khan, Shakila Thangaratinam.

Writing – original draft: John Allotey, Javier Zamora, Shakila Thangaratinam.

Writing – review & editing: John Allotey, Borja M. Fernandez-Felix, Javier Zamora, Ngawai Moss, Manny Bagary, Andrew Kelso, Rehan Khan, Joris A. M. van der Post, Ben W. Mol, Alexander M. Pirie, Dougall McCorry, Khalid S. Khan, Shakila Thangaratinam.

References

1. Edey S, Moran N, Nashef L. SUDEP and epilepsy-related mortality in pregnancy. Epilepsia. 2014; 55 (7):e72–4. https://doi.org/10.1111/epi.12621 PMID: 24754364
2. Knight M, Nair M, Tuffnell D, Kenyon S, Shakespeare J, Brocklehurst P, et al., editors. Saving lives, improving mothers' care—surveillance of maternal deaths in the UK 2012–14 and lessons learned to inform maternity care from the UK and Ireland Confidential Enquiries into Maternal Deaths and Morbidity 2009–14. Oxford: University of Oxford National Perinatal Epidemiology Unit; 2016.
3. Cantwell R, Clutton-Brock T, Cooper G, Dawson A, Drife J, Garrod D, et al. Saving mothers' lives: reviewing maternal deaths to make motherhood safer: 2006–2008. The Eighth Report of the Confidential Enquiries into Maternal Deaths in the United Kingdom. BJOG. 2011; 118 (Suppl 1):1–203. https://doi.org/10.1111/j.1471-0528.2010.02847.x PMID: 21356004
4. Fairgrieve SD. Population based, prospective study of the care of women with epilepsy in pregnancy. BMJ. 2000; 321(7262):674–5. https://doi.org/10.1136/bmj.321.7262.674 PMID: 10987772
5. Knight M, Kenyon S, Brocklehurst P, Neilson J, Shakespeare J, Kurinczuk JJ, editors. Saving lives, improving mothers' care: lessons learned to inform future maternity care from the UK and Ireland Confidential Enquiries into Maternal Deaths and Morbidity 2009–12. Oxford: University of Oxford National Perinatal Epidemiology Unit; 2014.
6. Man SL, Petersen I, Thompson M, Nazareth I. Antiepileptic drugs during pregnancy in primary care: a UK population based study. PLoS ONE. 2012; 7(12):e52339. https://doi.org/10.1371/journal.pone.0052339 PMID: 23272239
7. Nordeng H, Ystrom E, Einarson A. Perception of risk regarding the use of medications and other exposures during pregnancy. Eur J Clin Pharmacol. 2010; 66(2):207–14. https://doi.org/10.1007/s00228-009-0744-2 PMID: 19841915
8. Charyton C, Elliott JO, Lu B, Moore JL. The impact of social support on health related quality of life in persons with epilepsy. Epilepsy Behav. 2009; 16(4):640–5. https://doi.org/10.1016/j.yebeh.2009.09.011 PMID: 19854111
9. Loring DW, Meador KJ, Lee GP. Determinants of quality of life in epilepsy. Epilepsy Behav. 2004; 5(6):976–80. https://doi.org/10.1016/j.yebeh.2004.08.019 PMID: 15582847
10. Smeets VM, van Lierop BA, Vanhoutvin JP, Aldenkamp AP, Nijhuis FJ. Epilepsy and employment: literature review. Epilepsy Behav. 2007; 10(3):354–62. https://doi.org/10.1016/j.yebeh.2007.02.006 PMID: 17369102
11. Royal College of Obstetricians and Gynaecologists Epilepsy in pregnancy. Green-top Guideline No. 68. London: Royal College of Obstetricians and Gynaecologists; 2016 [cited 2019 Apr 16]. https://www.rcog.org.uk/globalassets/documents/guidelines/green-top-guidelines/gtg68_epilepsy.pdf.
12. Camfield P, Camfield C. Idiopathic generalized epilepsy with generalized tonic-clonic seizures (IGE-GTC): a population-based cohort with >20 year follow up for medical and social outcome. Epilepsy Behav. 2010; 18(1–2):61–3. https://doi.org/10.1016/j.yebeh.2010.02.014 PMID: 20471324
13. Edwards AG, Naik G, Ahmed H, Elwyn GJ, Pickles T, Hood K, et al. Personalised risk communication for informed decision making about taking screening tests. Cochrane Database Syst Rev. 2013;(2): CD001865. https://doi.org/10.1002/14651858.CD001865.pub3 PMID: 23450534
14. Battino D, Tomson T, Bonizzoni E, Craig J, Lindhout D, Sabers A, et al. Seizure control and treatment changes in pregnancy: observations from the EURAP epilepsy pregnancy registry. Epilepsia. 2013; 54 (9):1621–7. https://doi.org/10.1111/epi.12302 PMID: 23848605

15. Sveberg L, Svalheim S, Taubøll E. The impact of seizures on pregnancy and delivery. Seizure. 2015; 28:35–8. https://doi.org/10.1016/j.seizure.2015.02.020 PMID: 25746572

16. Thomas SV, Syam U, Devi JS. Predictors of seizures during pregnancy in women with epilepsy. Epilepsia. 2012; 53(5):e85–8. https://doi.org/10.1111/j.1528-1167.2012.03439.x PMID: 22429269

17. Thangaratinam S, Marlin N, Newton S, Weckesser A, Bagary M, Greenhill L, et al. AntiEpileptic drug Monitoring in PREgnancy (EMPiRE): a double-blind randomised trial on effectiveness and acceptability of monitoring strategies. Health Technol Assess. 2018; 22(23):1–152. https://doi.org/10.3310/hta22230 PMID: 29737274

18. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009; 338:b605. https://doi.org/10.1136/bmj.b605 PMID: 19477892

19. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009; 338:b375. https://doi.org/10.1136/bmj.b375 PMID: 19237405

20. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ. 2009; 338:b604. https://doi.org/10.1136/bmj.b604 PMID: 19336487

21. National Institute for Health and Care Excellence. Epilepsies, diagnosis and management. Clinical guideline CG137. London: National Institute for Clinical Excellence; 2018.

22. Harden CL, Hopp J, Ting TY, Pennell PB, French JA, Hauser WA, et al. Practice parameter update: management issues for women with epilepsy—focus on pregnancy (an evidence-based review): obstetrical complications and change in seizure frequency: report of the Quality Standards Subcommittee and Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology and American Epilepsy Society. Neurology. 2009; 73(2):126–32. https://doi.org/10.1212/WNL.0b013e3181a6b2f8 PMID: 19398682

23. Tomson T, Battino D, Bonizzoni E, Craig J, Lindhout D, Sabers A, et al. Dose-dependent risk of malformations with antiepileptic drugs: an analysis of data from the EURAP epilepsy and pregnancy registry. Lancet Neurol. 2011; 10(7):609–17. https://doi.org/10.1016/S1474-4422(11)70107-7 PMID: 21652013

24. Wilhelm J, Morris D, Hotham N. Epilepsy and pregnancy—a review of 98 pregnancies. The Aust N Z J Obstet Gynaecol. 1990; 30(4):290–5. https://doi.org/10.1111/j.1479-828X.1990.tb02013.x PMID: 2082882

25. Tomson T, Lindbom U, Ekqvist B, Sundqvist A. Epilepsy and pregnancy: a prospective study of seizure control in relation to free and total plasma concentrations of carbamazepine and phenytoin. Epilepsia. 1994; 35(1):122–30. https://doi.org/10.1111/j.1528-1157.1994.tb02921.x PMID: 8112234

26. Bardy AH. Incidence of seizures during pregnancy, labor and puerperium in epileptic women: a prospective study. Acta Neurol Scand. 1987; 75(5):356–60. https://doi.org/10.1111/j.1600-0404.1987.tb05459.x PMID: 3618113

27. Fisher RS, Cross JH, D'Souza C, French JA, Haut SR, Higurashi N, et al. Instruction manual for the ILAE 2017 operational classification of seizure types. Epilepsia. 2017; 58(4):531–42. https://doi.org/10.1111/epi.13671 PMID: 28276064

28. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996; 49(12):1373–9. https://doi.org/10.1016/s0895-4356(96)00236-3 PMID: 8970487

29. Moons KG, Altman DG, Reitsma JB. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015; 162(1): W1–73. https://doi.org/10.7326/M14-0698 PMID: 25560730

30. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd edition. New York: John Wiley & Sons; 2000.

31. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd edition. New York: John Wiley; 2002.

32. Royston P, Sauerbrei W. Building multivariable regression models with continuous covariates in clinical epidemiology—with an emphasis on fractional polynomials. Method Inform Med. 2005; 44(4):561–71.

33. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005; 58(5):475–83. https://doi.org/10.1016/j.jclinepi.2004.06.017 PMID: 15845334

34. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996; 15(4):361–87. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4 PMID: 8668867

35. 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(1):128–38. https://doi.org/10.1097/EDE.0b013e3181c30fb2 PMID: 20010215

36. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006; 26(6):565–74. https://doi.org/10.1177/0272989X06295361 PMID: 17099194

37. StataCorp. Stata statistical software. Release 15. College Station (TX): StataCorp; 2015.
38. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010; 33(1):1–22. PMID: 20808728
39. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.
40. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med. 2011; 30(4):377–99. https://doi.org/10.1002/sim.4067 PMID: 21225900
41. Aguglia U, Barboni G, Battino D, Cavazzuti GB, Citernesi A, Corosu R, et al. Italian consensus conference on epilepsy and pregnancy, labor and puerperium. Epilepsia. 2009; 50(Suppl 1):7–23. https://doi.org/10.1111/j.1528-1167.2008.01964.x PMID: 19125842
42. Thangaratinam S, Rikunenko R, Greenhill L, Bagary M, Pirie A, Khan KS, et al. Optimal monitoring of anti epileptic drugs in pregnancy: time for a randomised controlled trial? Arch Dis Child Fetal Neonatal Ed. 2011; 96(Suppl 1):Fa120. https://doi.org/10.1136/adc.2011.300163.79
43. Scottish Intercollegiate Guidelines Network. Diagnosis and management of epilepsy in adults. SIGN 143. Edinburgh: Scottish Intercollegiate Guidelines Network; 2018.
44. Harden CL, Pennell PB, Koppel BS, Hovinga CA, Gidal B, Meador KJ, et al. Practice parameter update: management issues for women with epilepsy—focus on pregnancy (an evidence-based review): vitamin K, folic acid, blood levels, and breastfeeding: report of the Quality Standards Subcommittee and Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology and American Epilepsy Society. Neurology. 2009; 73(2):142–9. https://doi.org/10.1212/WNL.0b013e3181a6b325 PMID: 19398680
45. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012; 98(9):683–90. https://doi.org/10.1136/heartjnl-2011-301246 PMID: 22397945
46. Bromley R, Weston J, Adab N, Greenhalgh J, Sanniti A, McKay AJ, et al. Treatment for epilepsy in pregnancy: neurodevelopmental outcomes in the child. Cochrane Database Syst Rev. 2014;(10): CD010236. https://doi.org/10.1002/14651858.CD010236.pub2 PMID: 25354543
47. World Health Organization. mhGAP: scaling up care for mental, neurological and substance use disorders. Geneva: World Health Organization; 2011 [cited 2019 Apr 16]. http://www.who.int/mental_health/mhgap_final_english.pdf.
48. Harden CL, Hopp J, Ting TY, Pennell PB, French JA, Allen Hauser W, et al. Management issues for women with epilepsy—focus on pregnancy (an evidence-based review): I. Obstetrical complications and change in seizure frequency: report of the Quality Standards Subcommittee and Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology and the American Epilepsy Society. Epilepsia. 2009; 50(5):1229–36. https://doi.org/10.1111/j.1528-1167.2009.02128.x PMID: 19496807
49. Vajda FJ, Hitchcock A, Graham J, O'Brien T, Lander C, Eadie M. Seizure control in antiepileptic drug-treated pregnancy. Epilepsia. 2008; 49(1):172–6. https://doi.org/10.1111/j.1528-1167.2007.01412.x PMID: 18031551
50. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016; 35(2):214–26. https://doi.org/10.1002/sim.6787 PMID: 26553135
51. Lamberink HJ, Otte WM, Geerts AT, Pavlovic M, Ramos-Lizana J, Marson AG, et al. Individualised prediction model of seizure recurrence and long-term outcomes after withdrawal of antiepileptic drugs in seizure-free patients: a systematic review and individual participant data meta-analysis. Lancet Neurol. 2017; 16(7):523–31. https://doi.org/10.1016/S1474-4422(17)30114-X PMID: 28483337
52. Kim LG, Johnson TL, Marson AG, Chadwick DW. Prediction of risk of seizure recurrence after a single seizure and early epilepsy: further results from the MESS trial. Lancet Neurol. 2006; 5(4):317–22. https://doi.org/10.1016/S1474-4422(06)70383-0 PMID: 16545748
53. Epilepsy Action. A critical time for epilepsy in England: a study of epilepsy service provision in England by Epilepsy Action. Yeadon (UK): Epilepsy Action; 2013 [cited 2019 Apr 16]. https://www.epilepsy.org.uk/sites/epilepsy/files/campaigns/ACT/Epilepsy%20Action%20-%20A%20Critical%20Time%20(2013).pdf.
54. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015; 68(3):279–89. https://doi.org/10.1016/j.jclinepi.2014.06.018 PMID: 25179855

ESTUDIO 2

PROGNOSTIC MODELS FOR MORTALITY AFTER CARDIAC SURGERY IN PATIENT WITH INFECTIVE ENDOCARDITIS: A SISTEMATIC REVIEW AND AGGREGATION OF PREDICTION MODELS

4.2. ESTUDIO 2: PROGNOSTIC MODELS FOR MORTALITY AFTER CARDIAC SURGERY IN PATIENTS WITH INFECTIVE ENDOCARDITIS: A SYSTEMATIC REVIEW AND AGGREGATION OF PREDICTION MODELS

4.2.1. RESUMEN

Hay ámbitos de la medicina en los que existen múltiples modelos pronósticos desarrollados con el objetivo de predecir un mismo resultado clínico de interés para una misma población diana. Así, en pacientes con endocarditis infecciosa, en los últimos años se han desarrollado varios modelos pronósticos para predecir el riesgo de mortalidad postoperatoria.

En este estudio se ha llevado a cabo una revisión sistemática y evaluación crítica de los modelos predictivos de mortalidad postoperatoria en pacientes sometidos a cirugía cardiaca con diagnóstico de endocarditis infecciosa. Tras la fase de revisión sistemática se procedió a la síntesis cuantitativa de los modelos encontrados para la derivación de un meta-modelo pronóstico.

Se realizó una búsqueda de los modelos pronósticos en las bases de datos de *Medline* y *EMBASE* desde su inicio hasta junio del 2020. Se incluyeron en la revisión sistemática los estudios que habían desarrollado o actualizado un modelo predictivo de mortalidad postoperatoria en pacientes diagnosticados de endocarditis infecciosa. La información relevante de cada modelo se obtuvo empleando un cuaderno de recogida de datos adaptado de la herramienta CHARMS (*CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies*) (64), y el riesgo de sesgo de los modelos se evaluó mediante la herramienta PROBAST (*Prediction model Risk Of Bias ASsessment Tool*) (65). Dos revisores, de forma independiente, realizaron el proceso de selección de estudios, extracción de datos y evaluación de la calidad de los modelos.

Los modelos incluidos en la revisión que fueron clasificados de bajo riesgo de sesgo fueron agregados en un meta-modelo basado en métodos de regresión apilada (en inglés: *stacked regression*), y los coeficientes del meta-modelo fueron optimizados con los datos de los pacientes incluidos en el registro nacional de endocarditis GAMES (Grupo de Apoyo al Manejo de la Endocarditis infecciosa en España) (90). Desde 2008 este registro incluye prospectivamente y de forma consecutiva y estandarizada todos los episodios de endocarditis infecciosa de 34 hospitales españoles. Para el presente estudio, seleccionamos todos los episodios infecciosos (n=1453) registrados en la cohorte GAMES de pacientes adultos sometidos a cirugía cardíaca

con diagnóstico preoperatorio de endocarditis infecciosa activa. El rendimiento predictivo del meta-modelo fue evaluado usando técnicas bootstrapping para obtener las medidas de rendimiento ajustadas por el potencial exceso de optimismo.

Se identificaron 11 modelos pronósticos de mortalidad postoperatoria en pacientes con endocarditis infecciosa. Ocho de estos modelos tenían alto riesgo de sesgo y fueron excluidos del proceso de agregación. El meta-modelo incluyó los predictores ponderados desde los tres modelos que fueron clasificados con riesgo de sesgo bajo o incierto (EndoSCORE, Specific ES-I y Specific ES-II) (91,92). Los coeficientes de los predictores edad y etiología de la infección, los cuales habían sido manejados de forma heterogénea en los estudios originales, fueron estimados a partir de los datos del registro GAMES.

El rendimiento predictivo del meta-modelo fue mejor que el de los modelos originales existentes, con las siguientes medidas de rendimiento: estadístico *C* 0.79 (IC 95% 0.76 a 0.82), *pendiente de calibración* 0.98 (IC 95% 0.86 a 1.13) y *calibration-in-the-large* -0.05 (IC 95% -0.20 a 0.11).

4.2.2. JUSTIFICACIÓN Y ASPECTOS METODOLÓGICOS

Son varios los aspectos metodológicos que merecen ser destacados en este segundo artículo. Estos se refieren a la utilidad del diseño de un protocolo del estudio y la metodología para la síntesis de los modelos predictivos.

El protocolo de esta revisión sistemática fue publicado en el registro PROSPERO y puede ser consultado por los lectores de la revisión (número de registro: CRD42020192602). En el apéndice de la tesis se incluye el documento completo.

En la introducción se ha mencionado que para la pregunta de investigación en estudios de pronóstico se debe emplear una versión extendida del acrónimo PICO, el acrónimo PICOTA (57). A pesar de que en nuestra revisión sistemática la pregunta de investigación fue formulada siguiendo este acrónimo, dada la naturaleza de la revisión nosotros no tuvimos un modelo pronóstico comparador o control. El interés radicó en identificar todos los modelos pronósticos sin una inclinación particular por alguno de ellos. En cuanto a la relación temporal entre la determinación de los predictores y la evaluación de los desenlaces (*timing*), se pueden considerar dos aspectos importantes: 1.) en el protocolo de la revisión se especifica que se incluirán modelos diseñados para predecir mortalidad temprana, independientemente de si la definición de los autores fue mortalidad en los primeros 30 días tras la operación o mortalidad

intrahospitalaria; 2.) dado que los modelos son desarrollados para ayudar en el manejo y la decisión sobre la conveniencia o no de la cirugía cardiaca, los predictores incluidos en los modelos deben estar medidos antes del inicio de la operación. A pesar de estas consideraciones, esto no fue motivo de exclusión a la hora de decidir si el modelo se incluía en la revisión o no. Aquellos modelos que consideraron algún predictor que debe ser medido durante o después de la cirugía fueron penalizados en la valoración de la aplicabilidad y no fueron considerados en el proceso de agregación. Respecto al ámbito de desarrollo y aplicación del modelo, en el protocolo se indicó que se considerarían todos los estudios independientemente del contexto, aunque se esperaba que todos hubieran sido desarrollados en un entorno hospitalario.

Con el objetivo de mejorar la efectividad para la identificación de los estudios de interés de nuestra revisión, en la estrategia de búsqueda. además del filtro metodológico propuesto por Geersing *et al.* (60), incluimos términos relacionados con la población objetivo, concretamente con la patología (endocarditis) y el tratamiento de elección (cirugía). Esto permitió reducir de forma considerable el número necesario de estudios a leer (NNR), magnitud que describe el número de referencias irrelevante identificadas en la búsqueda y que deben ser cribadas por cada estudio que es incluido en la revisión (61). A pesar de estos mecanismos de optimización de las búsquedas, se revisaron un total de 4.862 títulos para finalmente incluir en la revisión 9 estudios.

Se elaboró un formulario estandarizado (formato Excel) con una adaptación de la herramienta CHARMS (64) para la extracción de datos de los estudios y una adaptación de la herramienta PROBAST para evaluación del riesgo de sesgo (65). Se recogió la información relevante de los siguientes dominios: información general del estudio, fuentes de datos, participantes, desenlaces, candidatos predictores y métodos de análisis. De forma automatizada se importaban los ítems relevantes de cada dominio para facilitar la evaluación del riesgo de sesgo de los estudios y el diseño de tablas resumen.

A pesar de que una condición *sine qua non* para poder usar y validar externamente un modelo pronóstico es disponer de la ecuación completa del modelo (los coeficientes de los predictores y de la constante), tan solo cuatro (36%) de los once modelos que se identificaron en la búsqueda habían reportado la ecuación completa. Desafortunadamente, y pese a contactar con los autores de correspondencia de los estudios originales, solo pudimos validar externamente 6 modelos que facilitaron la ecuación completa del modelo.

El aspecto metodológico más destacado en este segundo estudio fue la metodología para la síntesis de los modelos pronósticos. La agregación de modelos es una estrategia atractiva cuando existen varios modelos predictivos en la literatura y se dispone de un conjunto de datos de validación. Esta estrategia consiste en combinar los modelos existentes en la literatura ponderando las asociaciones predictor-desenlace de los modelos originales acorde al rendimiento de cada modelo en la muestra de validación. En este estudio se ha seguido la metodología de regresiones apiladas para la agregación de los modelos (68,69,93). En esencia, la metodología consiste en incluir las predicciones de cada modelo de la literatura como una variable del meta-modelo y a continuación crear una combinación lineal de las predicciones del modelo.

En nuestra revisión sistemática se identificaron 9 estudios que informaron 11 modelos pronósticos, de los cuales sólo los tres que no presentaron alto riesgo de sesgo fueron considerados para la agregación.

Como en cualquier metaanálisis, la combinación del parámetro de interés, en nuestro estudio de los coeficientes de los modelos, solo es adecuada si las definiciones de los predictores son suficientemente homogéneas entre ellos. Cuando la agregación de los coeficientes de los predictores no fue posible, bien por la heterogeneidad en las definiciones, o bien porque su manejo fue distinto en los modelos publicados, se eliminaron del cálculo del predictor lineal y fueron estimados de forma independiente empleando los datos de la cohorte de validación. En concreto, la variable edad fue modelizada como continua en dos modelos y categorizada con múltiples puntos de corte en otro modelo y la variable etiología de la infección presentaba distintas categorías de agrupación entre los modelos incluidos. Por estos motivos los coeficientes para ambas variables fueron estimados en la muestra de validación sin tener en cuenta los coeficientes de los modelos publicados.

En la figura 7 se presenta de forma general las posibles casuísticas en la agregación de los coeficientes de los modelos para la generación del meta-modelo (68,69).

En el ejemplo se dispone de 3 modelos pronóstico: El primer modelo incluye seis predictores (*X1, X2, X4, X5, X7* y *X8*); el segundo modelo siete (*X1, X2, X3, X4, X6, X7* y *X8*); y el tercer modelo cinco (*X1, X2, X6, X7* y *X8*). Todos los modelos además incluyen el coeficiente de la constante (*Cons*).

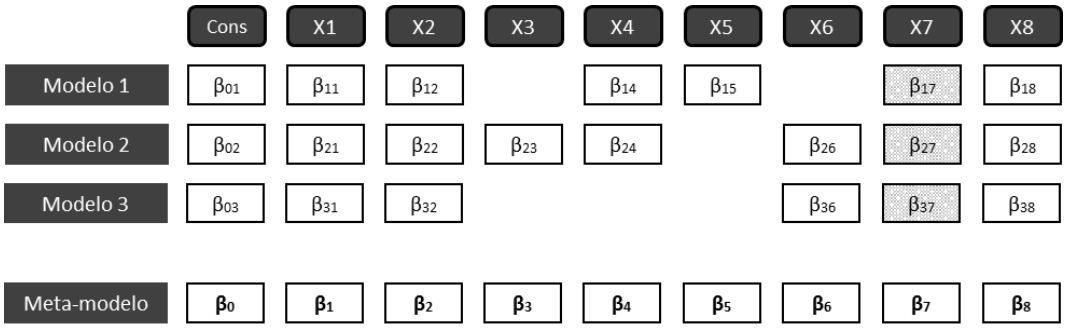


FIGURA 7. EJEMPLO GENERAL DE LAS POSIBLES CASUÍSTICAS EN LA AGREGACIÓN DE MODELOS PRONÓSTICOS

Los predictores *X1, X2, X7* y *X8* han sido incluidos en los tres modelos y los predictores *X3, X4, X5* y *X6* aparecen en al menos uno de los modelos. De los predictores comunes en todos los modelos *X1, X2* y *X8* presentan definiciones suficientemente homogéneas para poder ser combinados. Por su parte, el predictor *X7* (con relleno sólido gris) fue considerado en todos los modelos, pero presenta definiciones o formas de modelado demasiado heterogéneas, motivo que imposibilita su combinación en un coeficiente resumen. El resto de predictores no presentan divergencias en cuanto a su definición o modelado y pueden ser combinados. El coeficiente del predictor no incluido en el modelo se asume que toma valor cero en ese modelo.

El proceso de síntesis de los tres modelos de la figura 7 en un modelo agregado o meta-modelo se describe a continuación:

1. Para cada individuo de la muestra de validación se calculan los predictores lineales (LP) de los modelos existentes. Se excluyen del cálculo los predictores que no pueden ser combinados por la heterogeneidad presente como es el caso del predictor *X7* del ejemplo de la figura 7.

$$LP_m^{\dagger} = \beta_{m0} + \beta_{m1} \times X_1 + \beta_{m2} \times X_2 \ldots + \beta_{m8} \times X_8$$

ECUACIÓN 3. CÁLCULO DE LOS PREDICTORES LINEALES DE LOS MODELOS EXISTENTES

† Se excluye del cálculo el predictor X7.

2. Utilizando los datos de la muestra de validación se plantea un modelo de regresión logística con variable dependiente el desenlace de interés (mortalidad postoperatoria en nuestro meta-modelo) y como variables independientes los predictores lineales (LPs†) de cada modelo más los predictores excluidos en la ecuación 3 (el predictor X7).

$$ln\left(\frac{p_i}{1-p_i}\right) = \alpha_0 + \alpha_1 LP_1^{\dagger} + \alpha_2 LP_2^{\dagger} + \alpha_3 LP_3^{\dagger} + \alpha_4 X_7$$

ECUACIÓN 4. MODELO LOGÍSTICO CON LOS PREDICTORES LINEALES (LP†) DE LOS 3 MODELOS Y EL PREDICTOR X7 COMO VARIABLES INDEPENDIENTES

Las predicciones de cada modelo son ponderadas por un parámetro independiente α_j (que ha sido restringido a tomar valores ≥ 0 para que los modelos con una aportación negativa sean descartados del meta-modelo) que enfatiza modelos con una buena predicción global y penaliza modelos con pobre rendimiento o predicciones extremas en la muestra de validación (similar a la recalibración de los coeficientes o *logistic calibration* en los métodos de actualización de modelos) (49).

3. Los coeficientes α_1, α_2 y α_3 (ecuación 4) son los pesos por los que se ponderan los coeficientes de cada modelo en el meta-modelo agregado. El parámetro de peso residual α_0 (no es restringido) y asegura que el riesgo basal del meta-modelo sea óptimo en la muestra de validación (similar a la recalibración de la constante o *intercept updating* en los métodos de actualización de modelos) (49). El parámetro α_4 es el coeficiente del predictor X7 que ha sido estimado en la base de datos de validación.

4. Los coeficientes finales del meta-modelo para cada predictor son:

X1	$\beta_1 = \alpha_1 \times \beta_{11} + \alpha_2 \times \beta_{21} + \alpha_3 \times \beta_{31}$
X2	$\beta_2 = \alpha_1 \times \beta_{12} + \alpha_2 \times \beta_{22} + \alpha_3 \times \beta_{32}$
X3	$\beta_3 = \alpha_2 \times \beta_{23}$
X4	$\beta_4 = \alpha_1 \times \beta_{14} + \alpha_2 \times \beta_{24}$
X5	$\beta_5 = \alpha_1 \times \beta_{15}$
X6	$\beta_6 = \alpha_2 \times \beta_{26} + \alpha_3 \times \beta_{36}$
X7	$\beta_7 = \alpha_4$
X8	$\beta_8 = \alpha_1 \times \beta_{18} + \alpha_2 \times \beta_{28} + \alpha_3 \times \beta_{38}$

ECUACIÓN 5. ESTIMACIÓN DE LOS COEFICIENTES DEL META-MODELO

Al igual que en el modelo EMPiRE, con el fin de facilitar el uso del meta-modelo en la práctica clínica se ha desarrollado una calculadora online que permite a los clínicos conocer el riesgo de mortalidad post-operatoria del paciente antes de la cirugía basado en sus características. La calculadora está disponible en la plataforma Evidencio (https://www.evidencio.com/) y se puede acceder a ella a través de los filtros de búsqueda por especialidad clínica (enfermedades infecciosas) o indicando en el buscador "endocarditis".

4.2.3. APLICACIÓN

A continuación, se presenta un ejemplo del uso de la aplicación:

El meta-modelo es aplicado a una paciente de 60 años que presenta insuficiencia renal e hipertensión pulmonar, la fracción de eyección del ventrículo izquierdo parece normal (60%). La mujer no tiene enfermedad pulmonar crónica ni antecedentes de cirugía cardiaca. Su estado funcional según la escala NYHA es de tipo II y presenta complicaciones paravalvulares por la presencia de abscesos, pero sin presencia de fistulas. El estado preoperatorio no es crítico pero la paciente debe ser sometida urgentemente a la cirugía. El agente infeccioso fue *Staphylococcus spp.* y está localizado en la válvula aórtica. (Figura 8).

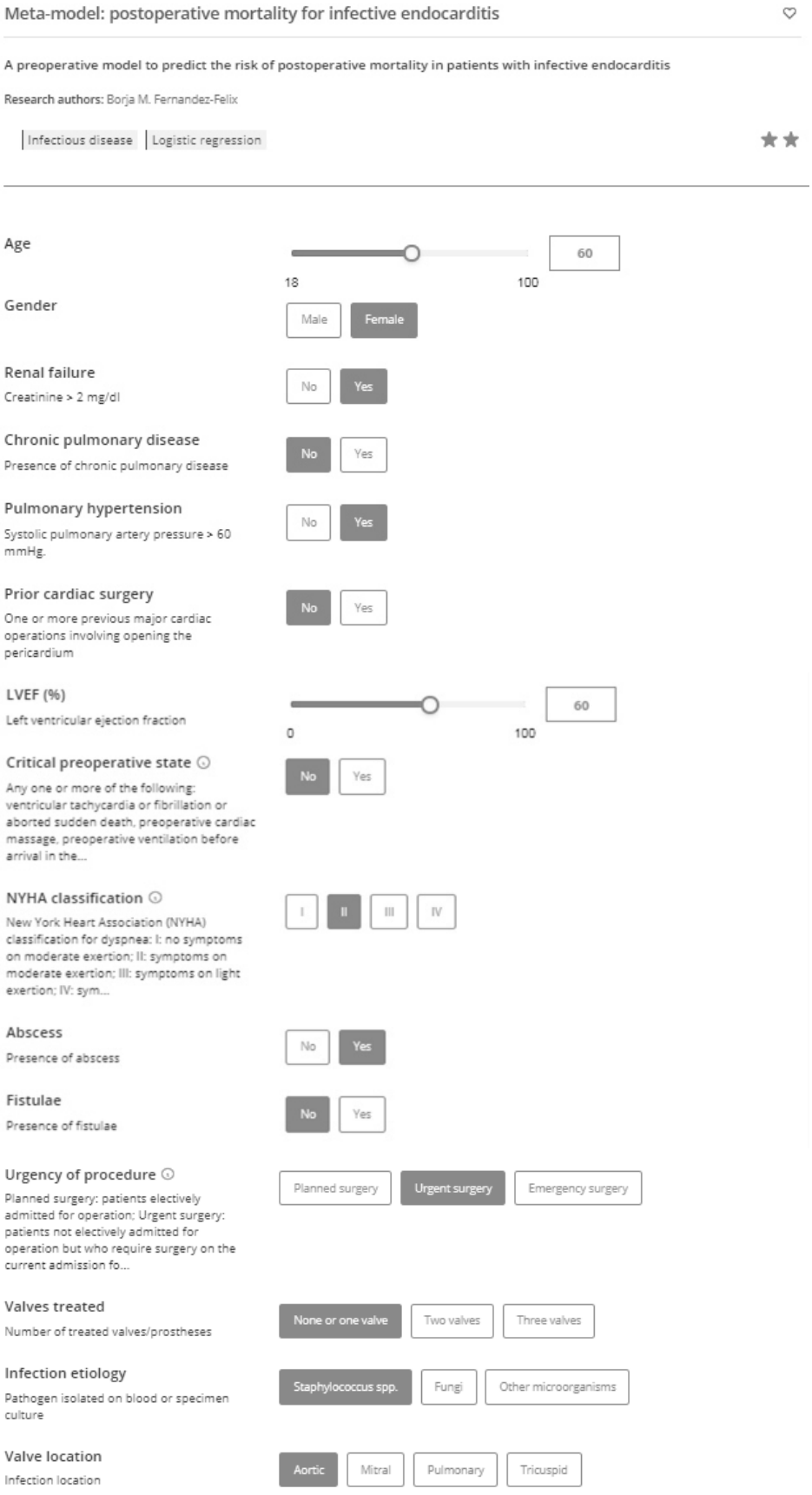


FIGURA 8. EJEMPLO DE APLICACIÓN DEL META-MODELO EN LA CALCULADORA ONLINE DE EVIDENCIO

La probabilidad de que esta paciente fallezca tras la cirugía cardiaca por la endocarditis infecciosa es del 37% (Figura 9).

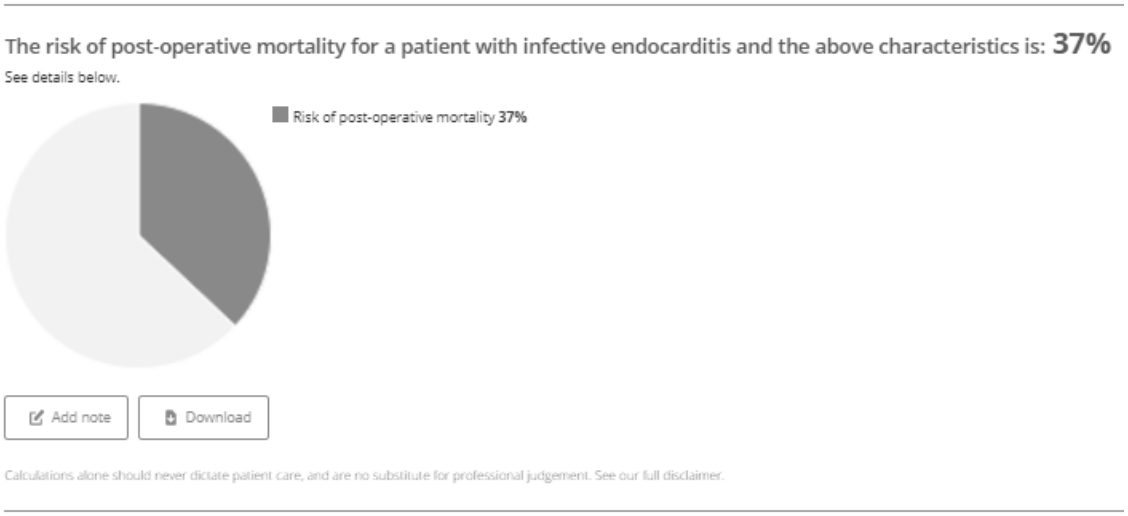


FIGURA 9. RIESGO DE MORTALIDAD POSTOPERATORIA ESTIMADO SEGÚN EL MODELO META-MODELO PARA LA PACIENTE DEL EJEMPLO

4.2.4. ARTÍCULO

Los resultados de este estudio han sido enviados con el título "*Prognostic models for mortality after cardiac surgery in patients with infective endocarditis: A systematic review and aggregation of prediction models*" a la revista Clinical Microbiology and Infection perteneciente al primer decil de la categoría "Infectious disease".

CMI
CLINICAL MICROBIOLOGY
AND INFECTION

ESCMID

Prognostic models for mortality after cardiac surgery in patients with infective endocarditis: a systematic review and aggregation of prediction models

Journal:	*Clinical Microbiology and Infection*
Manuscript ID	CLM-21-21010.R1
Article Type:	Systematic Review
Date Submitted by the Author:	n/a
Complete List of Authors:	Fernandez-Felix, Borja M; Hospital Universitario Ramon y Cajal, Clinical Biostatistics Unit. IRYCIS; CIBERESP, CIBER Epidemiology and Public Health Varela Barca, Laura; Fundación Jiménez Díaz, Department of Cardiovascular Surgery Garcia-Esquinas, Esther; CIBERESP, CIBER Epidemiology and Public Health; Universidad Autonoma de Madrid, Department of Preventive Medicine and Public Health. School of Medicine. IdiPaz Correa-Pérez, Andrea; Hospital Universitario Ramon y Cajal, Clinical Biostatistics Unit. IRYCIS; Universidad Francisco de Vitoria, Faculty of Medicine Fernández-Hidalgo, Nuria; Hospital Universitari Vall d'Hebron, Servei de Malalties Infeccioses; Instituto de Salud Carlos III, Red Española de Investigación en Patología Infecciosa (REIPI) Muriel, Alfonso; Universidad de Alcalá de Henares, Departamento de Enfermería y Fisioterapia; Hospital Universitario Ramon y Cajal, Clinical Biostatistics Unit. IRYCIS; CIBERESP, CIBER Epidemiology and Public Health Lopez-Alcalde, Jesus; Hospital Universitario Ramon y Cajal, Clinical Biostatistics Unit. IRYCIS; CIBERESP, CIBER Epidemiology and Public Health; Universidad Francisco de Vitoria, Faculty of Medicine; University Hospital Zurich and University of Zurich, Institute for Complementary and Integrative Medicine Álvarez-Díaz, Noelia; Hospital Universitario Ramon y Cajal, Medical Library Pijoan, Jose I; CIBERESP, CIBER Epidemiology and Public Health; Cruces University Hospital, OSI EEC; Biocruces-Bizkaia Health Research Institute Ribera, Aida; Hospital Universitari Vall d'Hebron, Cardiovascular Epidemiology and Research Unit; CIBERESP, CIBER Epidemiology and Public Health Navas, Enrique; Hospital Universitario Ramon y Cajal, Infectious Diseases Muñoz, Patricia; Hospital General Universitario Gregorio Marañón, Clinical Microbiology and Infectious Diseases ; Hospital General Universitario Gregorio Marañón-CIBERES, Clinical Microbiology and Infectious Diseases Fariñas, M Carnen; Hospital Universitario Marques de Valdecilla,

	Infectious Diseases Service. IDIVAL Goenaga, Miguel; Hospital Universitario de Donostia. IIS Biodonostia. OSI Donostialdea, Clinical Microbiology and Infectious Diseases Zamora , Javier ; Hospital Universitario Ramon y Cajal, Clinical Biostatistics Unit. IRYCIS; CIBERESP, CIBER Epidemiology and Public Health; University of Birmingham, WHO Collaborating Centre for Global Women's Health, Institute of Metabolism and Systems Research
Key Words:	Prognostic models, Systematic review, Meta-model, Aggregation, Validation, Infective Endocarditis
Abstract:	Background: There are several prognostic models to estimate the risk of mortality after surgery for active infective endocarditis (IE). However, these models incorporate different predictors and their performance is uncertain. Objective: We systematically reviewed and critically appraised all available prediction models of post-operative mortality in patients undergoing surgery for IE, and aggregated them into a meta-model. Data sources: We searched Medline and EMBASE databases from inception to June 2020. Study eligibility criteria: We included studies that developed or updated a prognostic model of post-operative mortality in patient with IE. Methods: We assessed the risk of bias of the models using PROBAST (Prediction model Risk Of Bias ASsessment Tool) and we aggregated them into an aggregate meta-model based on stacked regressions and optimized it for a nationwide registry of IE patients. The meta-model performance was assessed using bootstrap validation methods and adjusted for optimism. Results: We identified 11 prognostic models for post-operative mortality. Eight models had a high risk of bias. The meta-model included weighted predictors from the remaining three models (i.e., EndoSCORE, specific ES-I and specific ES-II), which were not rated as high risk of bias and provided full model equation. Additionally, two variables (i.e., age and infectious agent) which had been modelized differently across studies, were estimated based on the nationwide registry. The performance of the meta-model was better than the original three models, with the corresponding performance measures: C-statistics 0.79 (95% CI 0.76 to 0.82), calibration slope 0.98 (95% CI 0.86 to 1.13) and calibration-in-the-large -0.05 (95% CI -0.20 to 0.11). Conclusions: The meta-model outperformed published models and showed a robust predictive capacity for predicting the individualized risk of post-operative mortality in patients with IE. Protocol Registration: PROSPERO (registration number CRD42020192602)

Prognostic models for mortality after cardiac surgery in patients with infective endocarditis: a systematic review and aggregation of prediction models

Author list:

Borja M. Fernandez-Felix[1,2], Laura Varela Barca[3], Esther Garcia-Esquinas[2,4,5], Andrea Correa-Pérez[1,6], Nuria Fernández-Hidalgo[7,8], Alfonso Muriel[1,2], Jesus Lopez-Alcalde[1,2,6,9], Noelia Álvarez-Diaz[10], Jose I. Pijoan[2,11], Aida Ribera[2,12], Enrique Navas Elorza[13], Patricia Muñoz[14], Mª Carmen Fariñas[15], M. Ángel Goenaga[16], Javier Zamora[1,2,17]

Affiliations:

[1] Clinical Biostatistics Unit, Hospital Universitario Ramon y Cajal (IRYCIS), Madrid, Spain
[2] CIBER Epidemiology and Public Health (CIBERESP), Madrid, Spain
[3] Department of Cardiovascular Surgery Fundacion Jimenez Diaz University Hospital, Madrid, Spain
[4] Department of Preventive Medicine and Public Health. School of Medicine, Universidad Autónoma de Madrid, Madrid, Spain
[5] IdiPaz (Hospital Universitario La Paz-Universidad Autónoma de Madrid), Madrid, Spain
[6] Faculty of Medicine. Universidad Francisco de Vitoria, Madrid, Spain
[7] Servei de Malalties Infeccioses, Hospital Universitari Vall d'Hebron, Barcelona, Spain.
[8] Red Española de Investigación en Patología Infecciosa (REIPI), Instituto de Salud Carlos III, Madrid, Spain
[9] Institute for Complementary and Integrative Medicine, University Hospital Zurich and University of Zurich, Zurich, Switzerland
[10] Medical Library, Hospital Universitario Ramon y Cajal (IRYCIS), Madrid, Madrid, Spain
[11] Hospital Universitario Cruces/OSI EEC; Biocruces-Bizkaia Health Research Institute, Barakaldo, Spain
[12] Cardiovascular Epidemiology and Research Unit, Hospital Universitari Vall d'Hebron, Barcelona, Spain
[13] Department of Infectology, Hospital Universitario Ramon y Cajal (IRYCIS), Madrid, Spain
[14] Clinical Microbiology and Infectious Diseases Service, Hospital General Universitario Gregorio Marañón, Madrid. Instituto de Investigación Sanitaria Gregorio Marañón. CIBER Enfermedades Respiratorias-CIBERES. Facultad de Medicina, Universidad Complutense de Madrid, Spain
[15] Infectious Diseases Service. Hospital Universitario Marqués de Valdecilla-IDIVAL. Universidad de Cantabria, Santander, Spain.
[16] Infectious Diseases Service. Hospital Universitario Donostia. IIS Biodonostia. OSI Donostialdea. San Sebastián, Spain
[17] WHO Collaborating Centre for Global Women's Health, Institute of Metabolism and Systems Research, University of Birmingham, Birmingham, UK

Corresponding author:
Complete name: Borja Manuel Fernández Félix
Email address: borjamanuel86@gmail.com
ORCID: http://orcid.org/0000-0002-8798-019X
Postal address: Hospital Universitario Ramón y Cajal. Ctra. de Colmenar Viejo, Km. 9,100 (28034) MADRID
Phone number: +34 91 3368103
Fax number: +34 91 3369016

Category: Systematic review

Abstract

Background: There are several prognostic models to estimate the risk of mortality after surgery for active infective endocarditis (IE). However, these models incorporate different predictors and their performance is uncertain.

Objective: We systematically reviewed and critically appraised all available prediction models of post-operative mortality in patients undergoing surgery for IE, and aggregated them into a meta-model.

Data sources: We searched Medline and EMBASE databases from inception to June 2020.

Study eligibility criteria: We included studies that developed or updated a prognostic model of post-operative mortality in patient with IE.

Methods: We assessed the risk of bias of the models using PROBAST (Prediction model Risk Of Bias ASsessment Tool) and we aggregated them into an aggregate meta-model based on stacked regressions and optimized it for a nationwide registry of IE patients. The meta-model performance was assessed using bootstrap validation methods and adjusted for optimism.

Results: We identified 11 prognostic models for post-operative mortality. Eight models had a high risk of bias. The meta-model included weighted predictors from the remaining three models (*i.e.*, EndoSCORE, specific ES-I and specific ES-II), which were not rated as high risk of bias and provided full model equation. Additionally, two variables (*i.e.*, age and infectious agent) which had been modelized differently across studies, were estimated based on the nationwide registry. The performance of the meta-model was better than the original three models, with the corresponding performance measures: C-statistics 0.79 (95% CI 0.76 to 0.82), calibration slope 0.98 (95% CI 0.86 to 1.13) and calibration-in-the-large -0.05 (95% CI -0.20 to 0.11).

Conclusions: The meta-model outperformed published models and showed a robust predictive capacity for predicting the individualized risk of post-operative mortality in patients with IE.

Protocol Registration: PROSPERO (registration number CRD42020192602)

Key words: Prognostic models, systematic review, meta-model, aggregation, validation, infective endocarditis.

Background

Infective endocarditis (IE) is an uncommon but severe disease with a high mortality rate. Its current estimated incidence is 3-10 episodes per 100.000 person-years, while its in-hospital mortality rate ranges between 15% and 40% (1,2). Management of IE is often complex and, the decision whether to perform surgery remains a challenge because of the high mortality rate associated with the procedure. For that reason, it is estimated than less than half of the patients with surgical indication finally undergo cardiac surgery (3); which leads to a significantly decreased chance of survival (4). In this context, there has been a great interest in modeling prognosis of patients with IE to accurately estimate the risk of mortality in patients undergoing surgery for IE, and to help in the decision-making processes.

Prognostic models are mathematical equations that relates multiple variables for a particular individual to the probability of post-operative mortality. In the last decade, several IE prognostic models using preoperative patient´s-related and IE-specific factors, have been proposed. Unfortunately, these models have not been implemented in guidelines or are rarely applied in clinical practice. The poor adoption of these models could be a consequence of a shared perception of their limited validity because they have usually been built in relatively small cohorts and lack of external validation. Consequently, researchers carry on developing new models using their own data without considering prior knowledge, which leads to a scenario with multiple prognostic models of dubious validity. Therefore, we aimed to systematically review and critically appraise all available prediction models for post-operative mortality after cardiac surgery in patients with IE. We also aimed to aggregate those models with low risk of bias into a meta-model based on stacked regressions.

Methods

The protocol for this study was registered on PROSPERO (registration number CRD42020192602). We designed this systematic review according to the recent guidance (5,6), and reported its results following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) (7) and TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) recommendations (8,9).

Literature search

We searched Medline through Ovid and Embase through Elsevier from inception to 01/06/2020. We conducted a literature search to identify all potential studies for inclusion, without any language or publication dates restriction. We used the methodologic filter developed by Geersing et al. for prediction models research in MEDLINE (10), which was adapted for EMBASE. We added terms related to cardiac surgery and endocarditis. We further searched bibliographic references of included articles to identify other potential eligible studies. Complete search strings are shown in **Supplementary Material: S1.**

Eligibility criteria

We included original studies that developed prognostic models, with or without external validation, to predict the risk of post-operative mortality after cardiac surgery in patients with IE, as well as studies that updated previously published models. We accepted the authors` definition of post-operative mortality (either 30 days and/or in-hospital mortality), but excluded models that predicted mortality as part of a composite adverse outcome. Titles, abstracts, and full texts were screened for eligibility in pairs by three reviewers independently (BMFF, LVB, ACP) using EPPI-Reviewer 4 (11). Discrepancies were resolved by consensus.

Data extraction

Data extraction of included articles was done by three reviewers independently (pairs from BMFF, LVB, ACP). Discrepancies were solved by consensus. Reviewers used a standardized

data extraction form based on CHARMS (CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) (6). We extracted data on the following items: general information of the study, source of data, participants´ characteristics, outcome definition and time of occurrence, candidate predictors, and analysis methods. (**Supplementary Material: S2)**. When the completed model equation or relevant data were not provided, we contacted the correspondence authors to require this information.

Risk of bias assessment

We used a standardized form based on PROBAST (PRediction model risk of Bias ASsessment Tool) (12,13) to evaluate risk of bias (RoB) and applicability. We used the PROBAST definition of RoB. Concerns regarding the applicability of a primary study would arise when the population, predictors, or outcomes of the study differed from those specified in our review question. RoB and applicability were assessed by two independent reviewers (pairs from BMFF, LVB, ACP). We evaluated the relevant items on the following domains: Participants, predictors, outcome and analysis. Each domain was rated as a *high*, *low* or *unclear* RoB, and as providing *high*, *low* or *unclear* concerns regarding applicability. Any discrepancies were discussed between reviewers and resolved through discussion. The supplementary material provides details on critical appraisal and applicability (**Supplementary Material: S3**).

GAMES registry

We used the nationwide GAMES – Grupo de Apoyo al Manejo de la Endocarditis infecciosa en España – (14) registry as the validation dataset, to estimate existing models` weights for the meta-model development and its validation, and to externally validate the previously published models. Since January 2008, all consecutive episodes of IE in 34 Spanish hospitals were prospectively registered in GAMES using a standardized form. Regional and local ethics committees approved the study, and patients gave their informed consent in each center. For

the present study, we selected all the infective episodes (n=1,453) registered in the GAMES cohort involving adult patients (aged ≥18 years) who had undergone cardiac surgery with preoperative diagnosis of active IE. From these, 354 (24.4%) died after surgery (273 in the first 30 days and the remaining 81 during hospitalization). Assessment of predictors was done in an unblinded manner (i.e. with knowledge of the participant's outcome). **Supplementary Material: Table S1** shows the main descriptive characteristic of patients in the validation nationwide registry.

Statistical analyses

Model aggregation was based on stacked regressions (15). This methodology allows the synthesis of models collated in a systematic review into a meta-model using a validation dataset (16,17). We did not consider for aggregation the models that did not report the full equation or the models that were classified as high risk of bias. Stacked regressions used the linear predictor of each model as a co-variable in the meta-model, to subsequently created a linear combination of model predictions. That is, the original coefficients of each model are weighted by an independent parameter estimated in the meta-model, so that the models with worse performance in the validation dataset are penalized more. When aggregation of the coefficients was not possible, either because the definition of the predictor from primary studies was too heterogeneous or because predictors had been modeled differently in the published models (for instance, a numerical variable treated as a continuous predictor in one model and being categorized at different cut-points in the others), these predictors were dropped, and were included in the meta-model as independent covariables to re-estimate their coefficients entirely from scratch based on the validation dataset. Non-linear relationships for continuous predictors were tested using fractional polynomials (18).

Predictors with missing data in the validation dataset were imputed under the missing at random assumption using multiple imputation with chained equations (19). We included all

predictors and the outcome in the imputation models to ensure compatibility. (**Supplementary Material: S4**). Imputations checks were completed by looking at the distributions of imputed values to ensure plausibility. We generated 10 multiple imputed datasets and all primary analyses were performed in each imputed dataset. Pooled parameters were estimated both in the aggregation and validation processes using Rubin's rules (20).

The meta-model validation was assessed in terms of discrimination (*i.e.*, through the use of the C-statistic, with values from 1 indicating perfect discrimination to 0.5 no discrimination) and calibration (*i.e.*, through the calibration slope and calibration-in-the-large [CITL], with 1 and 0 as ideal values, respectively; as well as with calibration plots). Calibration plots represent the average predicted probability for risk groups categorized using deciles of predicted probability against observed proportion in each group, and fitting a lowess smoother to show calibration across the entire range of predicted probabilities at the individual-level (21,22). For the calibration plots we used the average predicted probabilities for individuals by pooling the imputed datasets using Rubin's rules (20). Because the meta-model was optimized to the validation dataset, we assessed its optimism-corrected performance measures by applying bootstrap validation with 500 replicates. As sensitivity analyses, we tested all model performance regardless of their critical appraisal. In addition, the meta-model performance was assessed only for 30-days mortality to investigate the meta-model robustness. To facilitate the use of the model, an online version of the prognostic tool was implemented in Evidencio (https://www.evidencio.com/). All analyses were performed using Stata software version 16 (23).

Results

Search results and study selection

We retrieved 4,862 titles through our systematic search combining Medline and Embase. From these, 684 duplicate references were identified. Of 4,178 titles assessed by title and

abstract, 34 studies were retained for full text screening, and 2 additional studies were detected in the bibliographic references of these articles. Nine studies describing 11 prediction models met the inclusion criteria (**Figure 1 and Supplementary Table S2**).

Source of data and participants

All included prognostic model development studies were published between 2011 and 2018. Six used data from a study cohort (three of them from a single center (24–26) and three from multiple centers (27–29)); two studies used data from multicenter registries (30,31); and one study used data from both a multicenter cohort and a local clinical registry (32). Eight studies used data from patients in Europe (Spain, Italy, France or Portugal) and one from patients in North America. Participants were recruited between 1980 and 2015. (**Supplementary Table S3**).

Outcomes

Three models were developed to predict any death occurring before discharge or within 30 days of surgery (24,26,30), five models to predict any death occurring before discharge (25,29,31,32), and the remaining three as death within 30 days of surgery (27,28). The incidence of deaths varied between 8.2% and 29.2% (**Table 1**).

Predictors

The number of candidate predictors considered in the models ranged from 15 to 57 and included patient-, clinical-, surgery- and IE-related factors. The number of parameters retained in the final models ranged from 2 to 15 (**Table 1**): The most common factors were critical preoperative state (n=9), renal failure (n = 7), age (n = 6), New York Heart Association (NYHA) classification of functional status (n=6), paravalvular complications (n = 6) and infection etiology (n = 5). The predictor definitions and the models' composition are shown in the **Supplementary Table S4 and Table S5**.

Model development and presentation

Sample sizes for models' development varied between 128 and 13,617 patients, and the number of events ranged from 21 to 1,117. Only two models from the same study adequately informed the handling of missing data (28), and these used complete data analyses. Logistic regression analysis was the most common modelling technique (n = 9), while logistic mixed effects (27) and logistic Generalized Estimating Equation (GEE) models (30) were only used in one model development each. Nine models used univariable analyses to select the candidate predictors. In nine out of eleven models the number of events per parameter (EPP) assessed for inclusion in the final model was lower than the minimum required for development of a new prediction model, based on the sample size estimation proposed by Riley et al.(33,34) (**Supplementary Table S6**). The method of predictors selection during multivariable modelling was backward selection in three models (25,32), stepwise selection in two models (29,31), and an automatic algorithm based on Akaike information criteria in multiple bootstrap samples in the other two models, with predictors selected in at least 70% of the bootstrapped samples being included in the final model (28). Four models did not inform about the method used to select predictors. (**Table 1**)

In seven out of 11 models the authors omitted the complete model equation (in five of them correspondence authors did not respond when were asked for further details) (**Supplementary Table S7**). Nine models were presented as a scoring system, and two of them included nomograms.

Model performance

The model performance was assessed in terms of discrimination through the C-statistic in all models. Nevertheless calibration was often wrongly assessed using the Hosmer-Lemeshow test (35) in six models. Only three models (26,28) used calibration slopes and CITL. Eight models were internally validated: three models were evaluated by bootstrapping with

correction for optimism (27,28), one was assessed through the 0.632 bootstrap method (25), two used temporal split samples (32) and two used random split samples (29,30). Three models only estimated the apparent performance (24,26,31). Three models were externally validated in the same development study using very small sample sizes, with only 18 events in the Olmos' model (29) and 21 in the Gatti's models (32). Clinical utility of the models was never assessed.

Risk of bias

The RoB was high in eight models, unclear in one (27) and low in the remaining two (28) (**Table 1, Supplementary Table S8 and Figure S1**). Two of the eight models with high RoB scored at "high risk" in the participants domain. Eight models scored at "high risk" in the analysis domain. Most of the models had small sample sizes and even the number of EPP was close to 1 in several models, increasing the risk of overfitting (34). Many studies decided model predictors based on univariable analysis, three reported only the apparent performance and two used random splitting validation. The calibration was sub-optimally assessed in all models classified as high risk of bias, with most of them using the Hosmer-Lemeshow test.

Derivation of the Meta-model

The eight models with high RoB were excluded from the statistical synthesis so that only the EndoScore, Specifics EuroSCORE-I (Specific ES-I) and EuroSCORE-II (Specific ES-II) models were aggregated in the meta-model. The model developed by Di Mauro (EndoSCORE) (27) included 15 parameters, while the other two (Specific ES-I and Specific ES-II) developed by Fernández-Hidalgo (28), presented 10 and 9 parameters respectively, from the EuroSCORE models predictors (36,37) and IE-related factors (**Table 2 and Supplementary Table S7**). The dependent variable for the meta-model was mortality (either 30-days or in-hospital).

To construct the meta-model, we first calculated the linear predictors (LP) from EndoSCORE, Specific ES-I and Specific ES-II for each observation in the validation dataset, after dropping the parameters for age and infection etiology because these variables were modelized differently in the different studies. Subsequently, we adjusted the meta-model using a logistic regression model, which incorporated the LPs as co-variables, to estimate the models' weights for aggregation, as well as the predictors for age (treated as continuous) and infection etiology (categorized into three groups: *Staphylococcus spp.*, fungi and other microorganisms) to re-estimate the coefficients from scratch. The meta-model included the predictors considered in at least one of the three original models. These are patient-related factors (i.e. age, gender, renal failure, prior cardiac surgery, chronic pulmonary disease, pulmonary hypertension and left ventricular ejection fraction), clinical presentation-related factors (i.e. critical preoperative state, New York Heart Association (NYHA) classification of functional status), surgery-related factors (i.e. presence of paravalvular complications (abscess and/or fistulae), urgency of procedure and number of treated valves/prostheses) and finally IE-related factors (i.e. valve location and infection etiology) (**Supplementary Table S5**). We have developed an online calculator to allow a simple and effective use of the meta-model. The magnitude of the associations of the predictive factors with mortality is shown in **Table 2** and the complete meta-model equation in **Supplementary Box S1**.

Validation of the models

The three prediction models considered for aggregation and the meta-model were validated in the GAMES registry. The C-statistics and their 95% confidence intervals (95%CI) for the published models were: 0.759 (95% CI 0.731 to 0.788) for EndoSCORE, 0.758 (95% CI 0.731 to 0.786) for Specific ES-I, and 0.762 (95% CI 0.735 to 0.789) for Specific ES-II. The optimism adjusted C-statistic for the meta-model was 0.79 (95% CI 0.76 to 0.82) (**Figure 2**). Calibration slopes were < 1 for all published models: 0.80 (95% CI 0.69 to 0.92) for

EndoScore, 0.82 (95% CI 0.70 to 0.94) for Specific ES-I, and 0.76 (95% CI 0.65 to 0.87) for Specific ES-II. CITL was 0.58 (95% CI 0.44 to 0.71) for EndoSCORE and 0.62 (95% CI 0.48 to 0.76) for Specific ES-II, and -0.02 (95% CI -0.16 to 0.11) for Specific ES-I. Optimism adjusted calibration measures for the meta-model were 0.98 (95% CI 0.86 to 1.13) for the slope and -0.05 (95% CI -0.20 to 0.11) for CITL (**Figure 2**). The calibration plots for the three previously published models and the meta-model are shown in **Figure 3**.

Sensitivity analysis showed that the meta-model had better overall performance than all published models regardless of their quality assessment (**Supplementary Figure S2**). Moreover, even though the meta-model was not fitted for the 30-days mortality outcome, it outperformed the three models used for model aggregation. (**Supplementary Figure S3**)

Discussion

Summary of findings

In this systematic review of prediction models for post-operative mortality in patients with infective endocarditis, we identified and critically appraised 11 models developed in 9 studies. The predicted outcome varied between studies (in-hospital, 30-days or both in-hospital or 30-days mortality). Of the eleven prognostic models, only two had low RoB and one unclear; the remaining eight models had high RoB mainly owing to poor statistical methods used, which suggests that their predictive performance when used in practice is probably lower than that reported. The sample sizes used to develop the models were limited and this is a well-known problem that leads to inaccurate predictions and consequently incorrect healthcare decisions in practice (34).

Four out of the 11 published models reported the full model equation required for a models' aggregation and a complete independent external validation as recommended by reporting guidelines (8,9). Two models' equations were recovered after request to the corresponding authors. Three models that were flagged as low or unclear RoB were aggregated to build the meta-model. Our meta-model included as predictors age, gender, renal failure, prior cardiac surgery, chronic pulmonary disease, pulmonary hypertension, left ventricular ejection fraction, critical preoperative state, New York Heart Association (NYHA) classification of functional status presence of paravalvular complications (abscess and/or fistulae), urgency of procedure, number of treated valves/prostheses, valve location and infection etiology. It showed better performance than the original models. We investigated the internal validity of the meta-model using bootstrap validation, and the results indicate there was no substantial over-optimism and that the validation sample was sufficiently large to combine and update the published models. Therefore, the meta-model is likely less prone to over-optimism and

more generalizable to new patient populations or settings, because it was built from the evidence of several patient cohorts and optimized to a nationwide registry.

Strengths and limitations

To our knowledge, this is the first systematic review with specific focus on prediction models of post-operative mortality in patients with infective endocarditis, with a thorough evaluation of the RoB, and using an external validation cohort to build a meta-model. We only combined the prediction models with low or unclear RoB and adjusted them to a new patient population. We used multiple imputation of predictors to avoid loss of useful information. The resulting meta-model incorporated prior knowledge optimally and outperformed previously published models.

Our study has some limitations. The outcome definition in the validation dataset was either 30-days or in-hospital post-operative mortality, and the outcome definition in the three models used for aggregation was 30-days mortality. Despite this difference a sensitivity analysis showed that the meta-model outperformed all published models when we explored its performance for the 30-days mortality. Two out of the three models considered for aggregation were developed in the same cohort. This circumstance increases the probability that the same predictors were included in both models and, therefore, it could magnify their associations with the outcome in the meta-model. However, we think that the impact of this magnification is limited because the weight of the ES-I model is relatively small compared to the other two models. Unfortunately, although we identified 11 prediction models in our systematic review, we were only able to validate the models for which the complete model equation was available. All these incomplete models were classified as high risk of bias and were consequently excluded from the analysis. We cannot rule out the presence of publication bias in our review. Unpublished studies are likely to be of poor quality (small, overfitted, and with poor predictive performance). Therefore, it is very likely that they would have been

excluded from our meta-model due to their high risk of bias. So the impact of this bias is expected to be low. Although the definition of predictors in GAMES registry was standardized, these could differ from definitions of published studies.

Comparison to existing studies

Most studies to develop new prediction models are based on small sample sizes and the modelling strategies are excessively driven by available data without considering the previous knowledge, leading to inefficient models. Other authors carried out external validation studies but none of them made a critical appraisal (38–41). In a previous study, Varela et. al. developed a prognostic model based on a systematic review of factors related to in-hospital mortality. The model was built using a series of univariate meta-analyses that pooled adjusted and unadjusted estimates altogether without taking into consideration the correlation among these factors. These pooled univariate estimates were then transformed into risk points to create a risk score (42,43). Our proposal includes more factors and our analysis included only estimates from low risk of bias studies. All estimates are from multivariate adjusted models and the weight each model has to build the meta-model is determined by their predictive performance in a validation cohort. This statistical methodology is in concordance with current recommendations (16,44).

Implications for practice

The decision whether to perform surgery for IE remains a challenge in clinical practice and it should come after a careful balance between the procedural risk and its estimated benefit. Critical preoperative state and priority of the procedure (urgent or emergency) are the most salient risk factors included in our meta-model. Patients with depressed LVEF, NYHA, renal failure have also worse prognosis. In addition, the aggressiveness of the IE infection as well as the technical difficulties of the surgery also implied higher risk of mortality. We expect a worse outcome in patients with IE caused by Staphylococcus or fungi or in patients with

paravalvular abscesses, fistulae or previous cardiac surgery because in these patients the surgery is challenging. Although risk scores for predicting mortality do not offer help in terms of establishing the burdens of surgical futility, they add a great value helping endocarditis teams to manage this complex disease and lead toward more personalized assistance based on individual patient characteristics. Moreover, the meta-model can be used to determine the case-mix of surgical hospitals and compare their performance adjusted for their case-mix.

Although in the 2015 IE guidelines (45) the score created by De Feo-Cotrufo et al for native IE is the only one recommended, it would be expected to change with the creation of several new IE specific scores and the generation of a meta-model that outperformed existing models.

The explanatory interpretation of the meta-model coefficients should be made with caution because coefficients have been shrunk, and therefore could be affected by the Stein's paradox (46). Shrinkage of the multivariable regression coefficients introduces a bias towards the null, but at the same time, properly shrinking coefficients ensures better predictions (47).

Challenges and opportunities

Further external validation studies are necessary to confirm the improvement in predictive ability of the meta-model. We will develop an online calculator to allow a simple and effective use of the meta-model. Given the low incidence of infective endocarditis, sufficiently large sample sizes for the adequate development of new predictive models are difficult to come by. We encourage authors to make their data available in order to allow building model based on available data (48,49).

Conclusions

The meta-model is a robust prognostic model to calculate the individualized risk of post-operative mortality in patients with infective endocarditis. It was developed based on the previous evidence using aggregation methods of the existing models identified from a

systematic review and after critical being appraised. The meta-model outperformed existing models; therefore, this preoperative tool can help guide individually tailored choices made by patients and clinicians.

Conflict of interest

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and the authors have declared that no competing interests exist.

Funding

CIBER (Biomedical Research Network in Epidemiology and Public Health) has partially supported the realization of this work (Grant number: ESP20G42X1). This public funding body had no role in the study design, the collection, analysis and interpretation of the data, the writing of the report nor the decision to submit the paper for publication.

Acknowledgements

We are grateful to the team involved with the collection of the GAMES registry data (**Supplementary Material: S6**).

Authors contributions

Conceptualization: BMFF, LVB, EGE, JLA, AM, JIP, AR, JZ; Search strategies: BMFF, NAD, JLA; Data extraction and Critical appraisal: BMFF, LVB, ACP; Methodology: BMFF, EGE, AM, JZ; Software, Formal analysis: BMFF; Validation: AM, JZ; Data adquisition/curation: BMFF, ENE, PM, MCF, MAG: Writing - Original draft: BMFF, EGE, JZ; Visualization: BMFF, LVB, NFH; Supervision: EGE, JZ; Writing – Review & Editing: All authors.

Bibliography

1. Murdoch DR. Clinical Presentation, Etiology, and Outcome of Infective Endocarditis in the 21st Century: The International Collaboration on Endocarditis–Prospective Cohort Study. Arch Intern Med. 2009 Mar 9;169(5):463.

2. Thuny F, Grisoli D, Collart F, Habib G, Raoult D. Management of infective endocarditis: challenges and perspectives. The Lancet. 2012 Mar;379(9819):965–75.

3. Iung B, Doco-Lecompte T, Chocron S, Strady C, Delahaye F, Le Moing V, et al. Cardiac surgery during the acute phase of infective endocarditis: discrepancies between European Society of Cardiology guidelines and practices. Eur Heart J. 2016 Mar 7;37(10):840–8.

4. Chu VH, Park LP, Athan E, Delahaye F, Freiberger T, Lamas C, et al. Association between surgical indications, operative risk, and clinical outcome in infective endocarditis: a prospective study from the International Collaboration on Endocarditis. Circulation. 2015 Jan 13;131(2):131–40.

5. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017 Jan 5;i6460.

6. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. PLoS Med. 2014 Oct 14;11(10):e1001744.

7. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009 Jul 21;6(7):e1000097.

8. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015 Jan 7;350:g7594.

9. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Annals of Internal Medicine. 2015 Jan 6;162(1):W1.

10. Geersing G-J, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons K. Search Filters for Finding Prognostic and Diagnostic Prediction Studies in Medline to Enhance Systematic Reviews. Smalheiser NR, editor. PLoS ONE. 2012 Feb 29;7(2):e32844.

11. Thomas, J., Graziosi, S., Brunton, J., Ghouze, Z., O'Driscoll, P., & Bond, M. (2020). EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis. EPPI-Centre Software. London: UCL Social Research Institute.

12. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med. 2019 Jan 1;170(1):W1.

13. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019 Jan 1;170(1):51.

14. Muñoz P, Kestler M, De Alarcon A, Miro JM, Bermejo J, Rodríguez-Abella H, et al. Current Epidemiology and Outcome of Infective Endocarditis: A Multicenter, Prospective, Cohort Study. Medicine (Baltimore). 2015 Oct;94(43):e1816.

15. Breiman L. Stacked regressions. Mach Learn. 1996 Jul;24(1):49–64.

16. Debray TPA, Koffijberg H, Nieboer D, Vergouwe Y, Steyerberg EW, Moons KGM. Meta-analysis and aggregation of multiple published prediction models: Meta-analysis and aggregation of multiple published prediction models. Statist Med. 2014 Jun 30;33(14):2341–62.

17. Martin GP, Mamas MA, Peek N, Buchan I, Sperrin M. A multiple-model generalisation of updating clinical prediction models. Statistics in Medicine. 2018 Apr 15;37(8):1343–58.

18. Royston P, Sauerbrei W. Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester, England ; Hoboken, NJ: John Wiley; 2008. 303 p. (Wiley series in probability and statistics).

19. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statist Med. 2011 Feb 20;30(4):377–99.

20. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.

21. Riley RD, Windt D van der, Croft P, Moons KGM. Prognosis research in healthcare: concepts, methods and impact. 2019.

22. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating [Internet]. 2019 [cited 2020 Apr 28]. Available from: https://doi.org/10.1007/978-3-030-16399-0

23. StataCorp. 2019. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC.

24. De Feo M, Cotrufo M, Carozza A, De Santo LS, Amendolara F, Giordano S, et al. The Need for a Specific Risk Prediction System in Native Valve Infective Endocarditis Surgery. The Scientific World Journal. 2012;2012:1–8.

25. Gatti G, Benussi B, Gripshi F, Della Mattia A, Proclemer A, Cannatà A, et al. A risk factor analysis for in-hospital mortality after surgery for infective endocarditis and a proposal of a new predictive scoring system. Infection. 2017 Aug;45(4):413–23.

26. Madeira S, Rodrigues R, Tralhão A, Santos M, Almeida C, Marques M, et al. Assessment of perioperative mortality risk in patients with infective endocarditis undergoing cardiac surgery: performance of the EuroSCORE I and II logistic models. Interact CardioVasc Thorac Surg. 2016 Feb;22(2):141–8.

27. Di Mauro M, Dato GMA, Barili F, Gelsomino S, Santè P, Corte AD, et al. A predictive model for early mortality after surgical treatment of heart valve or prosthesis infective endocarditis. The EndoSCORE. International Journal of Cardiology. 2017 Aug;241:97–102.

28. Fernández-Hidalgo N, Ferreria-González I, Marsal JR, Ribera A, Aznar ML, de Alarcón A, et al. A pragmatic approach for mortality prediction after surgery in infective endocarditis: optimizing and refining EuroSCORE. Clinical Microbiology and Infection. 2018 Oct;24(10):1102.e7-1102.e15.

29. Olmos C, Vilacosta I, Habib G, Maroto L, Fernández C, López J, et al. Risk score for cardiac surgery in active left-sided infective endocarditis. Heart. 2017 Sep;103(18):1435–42.

30. Gaca JG, Sheng S, Daneshmand MA, O'Brien S, Rankin JS, Brennan JM, et al. Outcomes for endocarditis surgery in North America: A simplified risk scoring system. The Journal of Thoracic and Cardiovascular Surgery. 2011 Jan;141(1):98-106.e2.

31. Martínez-Sellés M, Muñoz P, Arnáiz A, Moreno M, Gálvez J, Rodríguez-Roda J, et al. Valve surgery in active infective endocarditis: A simple score to predict in-hospital prognosis. International Journal of Cardiology. 2014 Jul;175(1):133–7.

32. Gatti G, Perrotti A, Obadia J, Duval X, Iung B, Alla F, et al. Simple Scoring System to Predict In-Hospital Mortality After Surgery for Infective Endocarditis. JAHA [Internet]. 2017 Jul [cited 2020 Dec 28];6(7). Available from: https://www.ahajournals.org/doi/10.1161/JAHA.116.004806

33. Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Statistics in Medicine. 2019 Mar 30;38(7):1276–96.

34. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020 Mar 18;m441.

35. On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative, Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019 Dec;17(1):230.

36. Nashef SAM, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). European Journal of Cardio-Thoracic Surgery. 1999 Jul;16(1):9–13.

37. Nashef SAM, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. EuroSCORE II. European Journal of Cardio-Thoracic Surgery. 2012 Apr 1;41(4):734–45.

38. Varela L, López-Menéndez J, Redondo A, Fajardo ER, Miguelena J, Centella T, et al. Mortality risk prediction in infective endocarditis surgery: reliability analysis of specific scores†. European Journal of Cardio-Thoracic Surgery. 2018 May 1;53(5):1049–54.

39. Pivatto Júnior F, Bellagamba CC de A, Pianca EG, Fernandes FS, Butzke M, Busato SB, et al. Análise de Escores de Risco para Predição de Mortalidade em Pacientes Submetidos à Cirurgia Cardíaca por Endocardite. ABC Cardiol [Internet]. 2020 [cited 2021 Mar 5]; Available from: https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0066-782X2020000300518

40. Gatti G, Sponga S, Peghin M, Givone F, Ferrara V, Benussi B, et al. Risk scores and surgery for infective endocarditis: in search of a good predictive score. Scandinavian Cardiovascular Journal. 2019 May 4;53(3):117–24.

41. Wang TKM, Oh T, Voss J, Gamble G, Kang N, Pemberton J. Comparison of contemporary risk scores for predicting outcomes after surgery for active infective endocarditis. Heart Vessels. 2015 Mar;30(2):227–34.

42. Varela Barca L, Navas Elorza E, Fernández-Hidalgo N, Moya Mur JL, Muriel García A, Fernández-Felix BM, et al. Prognostic factors of mortality after surgery in infective endocarditis: systematic review and meta-analysis. Infection. 2019 Dec;47(6):879–95.

43. Varela Barca L, Fernández-Felix BM, Navas Elorza E, Mestres CA, Muñoz P, Cuerpo-Caballero G, et al. Prognostic assessment of valvular surgery in active infective endocarditis: multicentric nationwide validation of a new score developed from a meta-analysis. European Journal of Cardio-Thoracic Surgery. 2020 Apr 1;57(4):724–31.

44. Debray TPA, Koffijberg H, Vergouwe Y, Moons KGM, Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. Statistics in Medicine. 2012 Oct 15;31(23):2697–712.

45. Habib G, Lancellotti P, Iung B. 2015 ESC Guidelines on the management of infective endocarditis: a big step forward for an old disease. Heart. 2016 Jul 1;102(13):992–4.

46. Efron B, Morris C. Stein's Paradox in Statistics. Scientific American. 1977;236(5):119–27.

47. van Houwelingen JC. Shrinkage and Penalized Likelihood as Methods to Improve Predictive Accuracy. Statistica Neerland. 2001 Mar;55(1):17–34.

48. Debray TPA, Riley RD, Rovers MM, Reitsma JB, Moons KGM, Cochrane IPD Meta-analysis Methods group. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. PLoS Med. 2015 Oct;12(10):e1001886.

49. Ensor J, Snell KIE, Debray TPA, Lambert PC, Look MP, Mamas MA, et al. Individual participant data meta-analysis for external validation, recalibration, and updating of a flexible parametric prognostic model. Stat Med. 2021 Jun 15;40(13):3066–84.

1 **Table 1. Models characteristics**

Author, Year Model name	Modelling method	Sample size	Events n (%)	Predictors		EPCP/ EPFP	Selection of candidate predictors	Selection of final predictors	Type of validation	Performance measures		Critical appraisal			
				Cand.	Final							P	Pr	O	A
In-hospital or 30 days mortality															
De Feo, 2012 [24] De Feo score	Logistic regression	440	40 (9.1)	19	6	2.1/ 6.7	Univariable (p-value < 0.05)	n.a.	Int: Apparent Ext: n.a.	Disc: C = 0.88 (0.82;0.93) Cal: HL Test	**RoB.**	–	?	+	–
											App.	–	+	+	
Gaca, 2011 [30] STS Score	Logistic GEE regression	13,617	1,117 (8.2)	38	13	29.4/ 85.9	Univariable and previous STS model variables	n.a.	Int: Random Split (D:70%/V:30%) Ext: n.a.	Disc: C = 0.76 Cal: Calibration plot	**RoB.**	–	+	+	–
											App.	+	+	+	
Madeira 2016 [26] -	Logistic regression	128	21 (16.4)	15	2	1.4/ 10.5	Univariable	n.a.	Int: Apparent Ext: n.a.	Disc: C = 0.87 (0.79;0.94) Cal: Slope; CITL	**RoB.**	?	+	+	–
											App.	?	+	+	
In-hospital mortality															
Gatti 2017a [32] AEPEI score	Logistic regression	361	56 (15.5)	57	5	1.0/ 11.2	Univariable (p-value < 0.1)	Backward	Int: 0.632 Bootstrap Ext: (n=161; e=21)	Disc: C = 0.72 (0.64;0.78) Cal: HL Test	**RoB.**	+	+	+	–
											App.	+	?	+	
Gatti 2017a [32] Alternate AEPEI score	Logistic regression	361	56 (15.5)	57	3	1.0/ 11.2	Univariable (p-value < 0.1)	Backward	Int: 0.632 Bootstrap Ext: (n=161; e=21)	Disc: C = 0.69 (0.61;0.76) Cal: HL Test	**RoB.**	+	+	+	–
											App.	+	+	+	
Gatti 2017b [25] ANCLA score	Logistic regression	138	28 (20.3)	56	5	0.5/ 5.6	Univariable (p-value < 0.1)	Backward	Int: 0.632 Bootstrap Ext: n.a.	Disc: C = 0.83 (0.75;0.89) Cal: HL Test	**RoB.**	+	+	+	–
											App.	+	+	+	
Martínez-Sellés 2014 [31] PALSUSE	Logistic regression	437	106 (24.3)	n.a.	7	n.a./ 15.1	Univariable (p-value < 0.1)	Stepwise	Int: Apparent Ext: n.a.	Disc: C = 0.84 (0.79;0.88) Cal: HL Test	**RoB.**	+	+	+	–
											App.	+	+	+	
Olmos 2017 [29] RISK-E	Logistic regression	424	124 (29.2)	37	8	3.4/ 15.5	Univariable (p-value < 0.1) and clinically relevant	Stepwise	Int: Random Split (D:66%/V:33%) Ext: (n=204; e=18)	Disc: C = 0.76 (0.64;0.88) Cal: HL Test; Calibration plot	**RoB.**	+	+	+	–
											App.	+	+	+	
30 days mortality															
Di Mauro 2017 [27] EndoSCORE	Logistic mixed effect regression	2,715	298 (11.0)	32	15	9.3/ 19.9	Univariable (p-value < 0.2)	n.a.	Internal: Bootstrap External: n.a.	Disc: C = 0.85 (0.84;0.86) Cal: CITL and slope vs. the ideal values	**RoB.**	?	+	+	?
											App.	?	+	+	
Fernández-Hidalgo 2018 [28] Specific ES-I	Logistic regression	779	208 (26.7)	26	10	8.0/ 20.8	Variables in ES-I and specific IE risk factor	Bootstrap	Int: Bootstrap Ext: n.a.	Disc: C = 0.77 (0.74;0.81) Cal: Slope = 0.93 CITL = -0.06	**RoB.**	+	+	+	+
											App.	+	?	+	
Fernández-Hidalgo 2018 [28] Specific ES-II	Logistic regression	779	208 (26.7)	27	9	7.7/ 23.1	Variables in ES-II and specific IE risk factor	Bootstrap	Int: Bootstrap Ext: n.a.	Disc: C = 0.77 (0.73;0.81) Cal: Slope = 0.93 CITL = -0.05	**RoB.**	+	+	+	+
											App.	+	+	+	

STS: Society of Thoracic Surgeons; AEPEI: Association pour l'Etude et la Prevention de l'Endocadite Infectieuse; ANCLA: Anemia, NYHA class IV, critical state, large intracardiac destruction, and surgery on thoracic aorta; PALSUSE: prosthetic valve, age≥70, large intracardiac destruction, Staphylococcus spp, urgent surgery, sex [female], EuroSCORE≥10; RISK-E: Risk-Endocarditis; ES: EuroSCORE; GEE: Generalized Estimating Equation; n: number of events; Cand: number of candidate predictors assessed: EPCP: events per candidate predictor; EPFP: events per final predictor; Critical appraisal domains (P: participants; Pr: predictors; O: outcome; A: analysis); n.a.: not available; Int: Internal validation (D: development cohort; V: validation cohort); Ext: external validation (n: sample size; e: number of events); Disc: Discrimination; Cal: calibration; HL: Hosmer-Lemeshow; CITL: calibration-in-the-large; RoB: Risk of Bias; App: applicability. +: Low RoB or low concern for applicability; –: High RoB or high concern for applicability; ?: Unclear RoB or applicability.

Table 2. Coefficients and odds ratios of the meta-model and the prediction models used for aggregation.

Predictors	Original models			Aggregated model	
	EndoSCORE Di Mauro 2017	Sp. ES-I Fernández-Hidalgo 2018	Sp. ES-II Fernández-Hidalgo 2018	Meta-model[a]	
				Coefficient (95% CI)	OR (95% CI)
Intercept	-2.60	-3.13	-4.21	-5.00 (-5.97; -4.00)	-
Gender (Female)	0.51			0.22 (0.14; 0.31)	1.25 (1.15; 1.36)
Age[b] (years)	-	-	-	0.045 (0.03; 0.06)	1.05 (1.03; 1.06)
Renal failure	0.50	0.46		0.28 (0.17; 0.41)	1.32 (1.19; 1.51)
Prior cardiac surgery		1.10	0.96	0.51 (0.36; 0.69)	1.67 (1.43; 1.99)
Chronic pulmonary disease	0.68			0.29 (0.19; 0.41)	1.34 (1.21; 1.51)
Pulmonary hypertension		1.27		0.17 (-0.11; 0.48)	1.19 (0.90; 1.62)
LVEF (%)	-0.03			-0.013 (-0.02; -0.01)	0.99 (0.98; 0.99)
Critical preoperative state	1.46	1.12	1.02	1.17 (0.97; 1.40)	3.22 (2.64; 4.06)
NYHA class. (>I)		0.70	0.62	0.33 (0.23; 0.44)	1.39 (1.26; 1.55)
Abscess	1.09			0.47 (0.30; 0.65)	1.60 (1.35; 1.92)
Fistulae		1.22	1.14	0.59 (0.42; 0.79)	1.80 (1.52; 2.20)
Priority of procedure					
- Urgent status			1.16	0.44 (0.16; 0.68)	1.55 (1.17; 1.97)
- Emergency status		0.81	1.95	0.85 (0.53; 1.17)	2.34 (1.70; 3.22)
Number of valves treated					
- Two valves treated	0.50			0.22 (0.14; 0.30)	1.25 (1.15; 1.35)
- Three valves treated	1.50			0.65 (0.41; 0.90)	1.92 (1.51; 2.46)
Valve location (Mitral)		0.37	0.38	0.19 (0.14; 0.25)	1.21 (1.15; 1.28)
Etiology[c]	-	-	-		
- *Staphylococcus* spp.				0.64 (0.35; 0.94)	1.90 (1.42; 2.56)
- Fungi				0.61 (-0.46; 1.40)	1.84 (0.63; 4.06)

LVEF: left ventricular ejection fraction; NYHA class: New York Health Association classification of functional status; OR: Odds ratio

[a] Weights used to create the meta-model: EndoScore = 0.433; Sp. ES-I = 0.131; Sp. ES-II = 0.379
Stacked regression:

$$\ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = -1.861 + 0.433 \times \mathrm{LP}^{\dagger}_{\mathrm{DM}} + 0.131 \times \mathrm{LP}^{\dagger}_{\mathrm{FH-I}} + 0.379 \times \mathrm{LP}^{\dagger}_{\mathrm{FH-II}} + 0.045 \times \mathrm{Age} + 0.64 \times \textit{Staphylococcus}\ \mathrm{spp.} + 0.61 \times \mathrm{Fungi}$$

Where, p is the probability of post-operative mortality and $LP_i^{\dagger}$ is the linear predictor for each model selected for aggregation dropping the parameters from age and infection etiology; DM (Di Mauro model [EndoSCORE]); FH-I (Fernández-Hidalgo model [sp. ES-I]); FH-II (Fernández-Hidalgo model [sp. ES-II]). Consequently, stacked intercept = -1.861 + 0.433 x (-2.60) + 0.131 x (-3.13) + 0.379 x (-4.21) = -5.00, and for instance, the stacked coefficient for renal failure = 0.433 x (0.50) + 0.131 x (0.46) + 0.379 x (0) = 0.277

[b] Age was categorized in Di Mauro 2017 and treated as continuous in Fernández-Hidalgo 2018
[c] Etiology was categorized in different ways in each existing model.

Figure 1. PRISMA flowchart of study inclusions and exclusions.

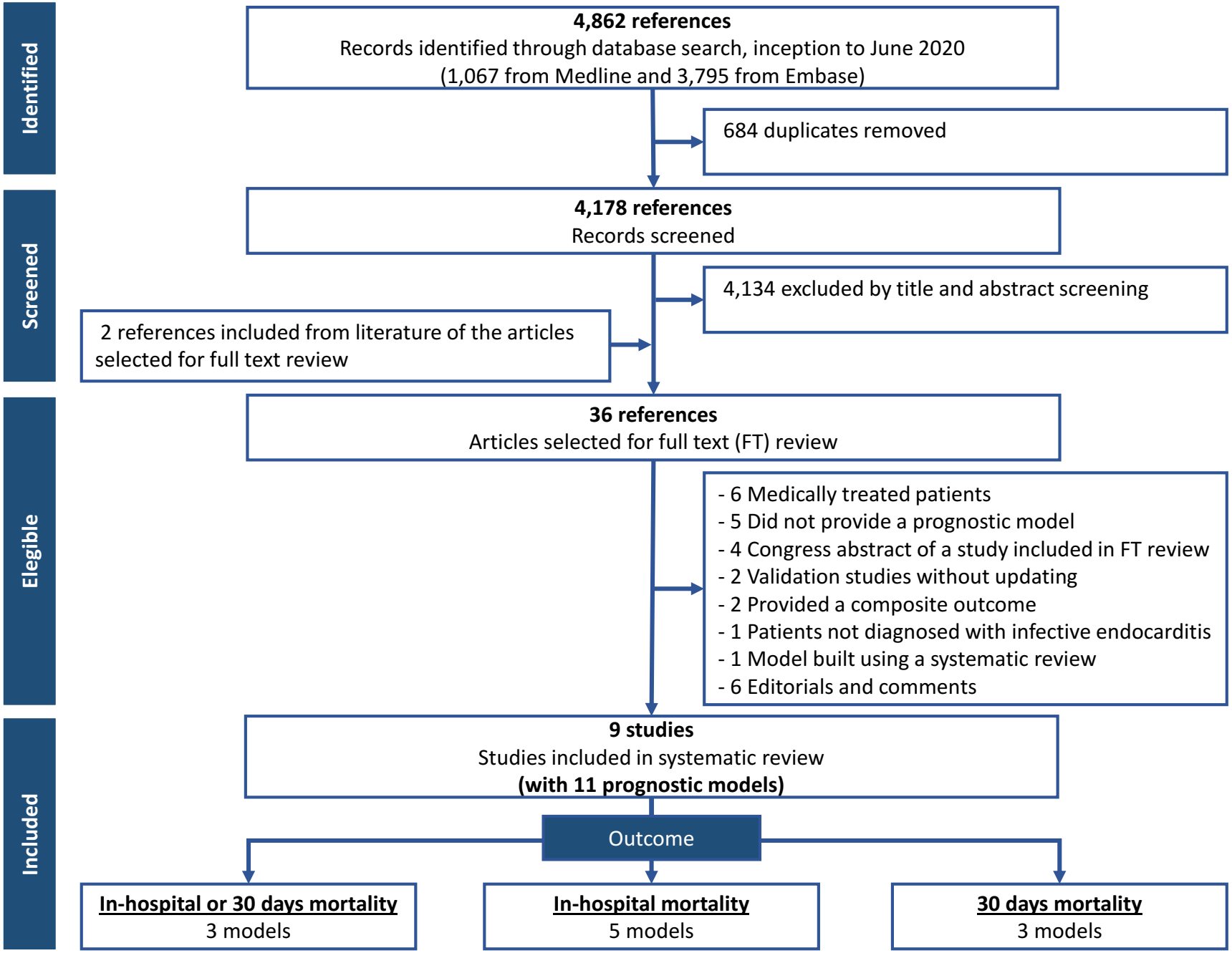
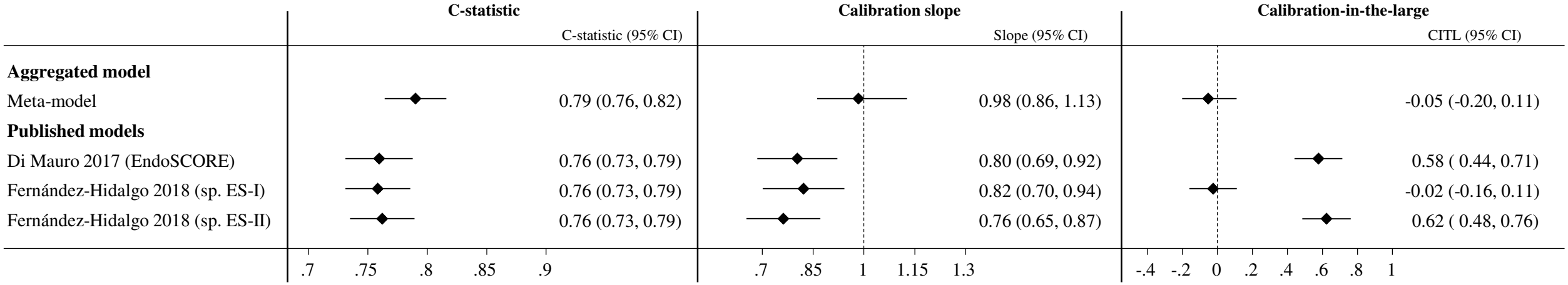
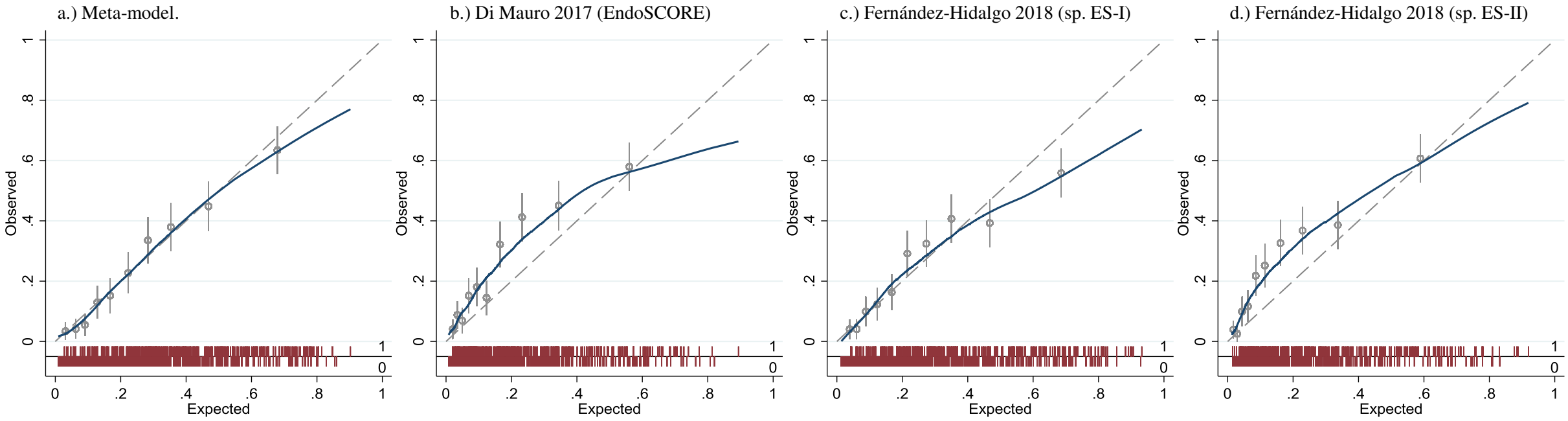


Figure 2. Bootstrap internal validation of the meta-model and external validation of existing models selected for aggregation



Dashed lines indicate lines of perfect calibration slope (1) and calibration-in-the-large (0). Black diamonds indicate point estimates and horizontal lines indicate 95% CIs. CITL: Calibration-in-the-large

Figure 3. Calibration plots of the meta-model and of the prediction models selected for aggregation.



Dashed lines represent perfect calibration, grey circles and bars indicate average risks and their confidence interval by deciles of the risk spectrum, dark blue lines indicate the lowess smoother assessment of the calibration at the individual level, and red spike plots show the distribution of events and non-events.

ESTUDIO 3

BOOTSTRAP INTERNAL VALIDATION COMMAND FOR PREDICTIVE LOGISTIC REGRESSION MODELS

BSVALIDATION

4.3. ESTUDIO 3: BOOTSTRAP INTERNAL VALIDATION COMMAND FOR PREDICTIVE LOGISTIC REGRESSION MODELS

4.3.1. RESUMEN

Cuando se desarrolla un modelo pronóstico el objetivo es predecir de forma fiable el desenlace de interés en futuros sujetos. El sobreajuste es un problema frecuente en el desarrollo de modelos pronósticos que amenaza la validez de las predicciones. Este problema supone que se obtienen estimaciones del rendimiento predictivo del modelo demasiado optimistas. La validación interna mediante técnicas de remuestreo bootstrapping permite cuantificar el exceso de optimismo del modelo. Ajustarlo y así reportar unas medidas del rendimiento más realistas.

En este estudio se ha desarrollado un comando de análisis para el software estadístico Stata (94). El comando – `bsvalidation` – permite llevar a cabo la validación interna de modelos de regresión logística binaria usando técnicas bootstrapping tras ejecutar el modelo final usando los comandos – `logistic` – o – `logit` – de Stata.

El comando calcula las medidas del rendimiento global, así como del rendimiento en términos de discriminación y calibración (ver Tabla 1). Además, permite generar gráficos de calibración usando una función de suavizado de los riesgos predichos y observados, y por grupos definidos por el usuario según los cuantiles de riesgo pronosticado. Adicionalmente, el programa calcula los factores de penalización (en inglés: *shrinkage factors*) que pueden usarse para ajustar uniformemente los coeficientes del modelo en presencia de sobreajuste.

El comando – `bsvalidation` – es una herramienta útil para realizar la validación interna de modelos pronóstico de regresión logística. Este hace más accesibles los métodos de validación interna bootstrap a la comunidad investigadora promoviendo un reporte más completo y mejorado de los estudios de desarrollo de modelos pronóstico.

4.3.2. JUSTIFICACIÓN Y ASPECTOS METODOLÓGICOS

A pesar de que el software estadístico Stata incorpora varios comandos post-estimación que permiten estimar las medidas del rendimiento aparente del modelo pronóstico, hasta nuestro conocimiento no existe ningún comando que permita realizar la validación interna mediante técnicas bootstrap y obtener el rendimiento del modelo corregido por el exceso de optimismo.

Son varios los aspectos metodológicos a destacar en este artículo. En primer lugar, subrayar la sencillez con la que se puede emplear la herramienta de validación desarrollada. Se trata de un comando post-estimación por lo que el usuario debe en primer lugar estimar el modelo logístico final y a continuación proceder con la validación.

En la tabla 1 se presentan las medidas del rendimiento que son estimadas por el comando – `bsvalidation` – y el rango de valores que puede tomar o su valor ideal y su interpretación.

Medida	Valores	Interpretación
Rendimiento global		
Brier scaled	Rango: [0 - 100]	Cuantifica el ajuste de las predicciones con el resultado real, en relación con la cantidad de variabilidad que se explica.
Discriminación		
C-Statistic	Rango: [0.5 - 1]	Cuantifica la capacidad del modelo para distinguir entre aquellos con y sin el resultado.
Calibración		
E:O ratio	Valor ideal: 1	**E:O ratio < 1:** indica que el modelo infraestima el número de eventos totales. **E:O ratio > 1:** indica que el modelo sobreestima el número de eventos totales.
Calibration-in-the-large (CITL)	Valor ideal: 0	**CITL < 0:** indica que las predicciones del modelo son sistemáticamente demasiado altas. **CITL > 0**: indica que las predicciones del modelo son sistemáticamente demasiado bajas
Calibration slope	Valor ideal: 1	**Slope < 1:** indica que las predicciones son demasiado extremas. El modelo está sobreajustado. **Slope > 1:** indica que las predicciones no varían lo suficiente. El modelo está infraajustadoe.

TABLA 1. MEDIDAS DE EVALUACIÓN DEL RENDIMIENTO DEL MODELO PRONÓSTICO

La salida de resultados del comando – `bsvalidation` – presenta la estimación de las medidas del rendimiento predictivo que se muestran en la tabla 1, tanto en términos de rendimiento aparente como corregidas por el exceso de optimismo.

Para evaluar la calibración del modelo se ha evitado el uso de la prueba estadística de *Hosmer-Lemeshow* (33). A pesar de que esta prueba es utilizada frecuentemente por la comunidad cientifica, sus resultados son muy sensibles al tamaño de la muestra y a la agrupación de los individuos en grupos de riesgo. Por ello las recomendaciones actuales desaconsejan su uso (9,10,28,29).

Un problema adicional en el desarrollo de modelos pronósticos es el sobreajuste. Este problema se produce cuando el modelo es demasiado complejo para la cantidad de información disponible. Un modelo sobreajustado describe muy bien la muestra de sujetos que han sido utilizados para su desarrollo, pero sus predicciones no serán tan acertadas cuando se aplica a nuevos sujetos. El problema de sobreajuste viene derivado de dos fuentes: la incertidumbre en la estimación de los parámetros del modelo y la incertidumbre propia del modelo. Es frecuente decidir las características que van a ser incluidas en el modelo como predictores en base a la información de los datos disponibles, generando incertidumbre en la especificación del modelo (19,20).

En modelos de predicción se puede definir el sobreajuste como el ajuste de un modelo estadístico con demasiados grados efectivos de libertad en el proceso de modelado (19). Los grados de libertad efectivos no sólo implican aquellos empleados en la estimación de los coeficientes del modelo final, sino también los empleados en todos los procesos del desarrollo desde la selección de predictores hasta las potenciales transformaciones no lineales de los predictores continuos. El sobreajuste es bastante frecuente en estudios de desarrollo de modelos predictivos, y dirige a estimaciones del rendimiento del modelo demasiado exageradas. Es decir, el investigador tiene la impresión de que el modelo funciona mejor de lo que funcionará cuando sea empleado en otros sujetos. Una alternativa para evitar el problema de sobreajuste es reducir los coeficientes de los predictores (19,20,95).

Es bien sabido, que los métodos de estimación basados en la función de máxima verosimilitud dirigen estimaciones insesgadas de los coeficientes, minimizando el error cuadrático medio entre los resultados predichos y observados. Sin embargo, la combinación de las estimaciones insesgada de los coeficientes no siempre es el mejor método para minimizar el error cuadrático medio del riesgo promedio predicho. Esto es conocido como la paradoja de Stein (96). El

concepto clave de esta paradoja es que se pueden obtener mejores predicciones aplicando un sesgo a los coeficientes del modelo a través de un factor de penalización o contracción (en inglés *shrinkage factor* (S)). Al aplicar un factor de penalización a los coeficientes estimados estos son contraídos en una regresión hacia la media.

Un factor de contracción uniforme para todos los predictores del modelo puede obtenerse mediante técnicas bootstrap como la diferencia entre la pendiente de calibración y el exceso de optimismo (19,20). Otra opción es el factor de contracción heurístico propuesto por *Van Houwelingen y Le Cessie* (44).

$$\widehat{S} = \frac{model\ X^2 - p}{model\ X^2}$$

ECUACIÓN 6. FACTOR DE CONTRACCIÓN HEURÍSTICO

Donde p es el número total de grados de libertad para los predictores y $model\ X^2$ es el estadístico X^2 de la razón de verosimilitud para testar la influencia conjunta de todos los predictores simultáneamente en el modelo completo.

Cuando se aplica un factor de contracción (S) a los coeficientes originales del modelo, la constante α^* es reestimada para asegurar la *calibration-in-the-large*.

$$ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

ECUACIÓN 7. ECUACIÓN DE UN MODELO CON LOS COEFICIENTES ORIGINALES

$$ln\left(\frac{p_i}{1-p_i}\right) = \alpha^* + S(\beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})$$

ECUACIÓN 8. ECUACIÓN DE UN MODELO CON LOS COEFICIENTES AJUSTADOS

El comando – `bsvalidation` – estima los factores de contracción bootstrap y heurístico, e incorpora la opción *adjust()* que permite al usuario ajustar los coeficientes originales del modelo por el factor de penalización elegido (*bootstrap* o *heurístic*), reajustando automáticamente la constante del modelo para asegurar la *calibration-in-the-large* (28).

Una forma eficiente de reducir el riesgo de sobreajuste del modelo es preespecificando la estructura de este. Es decir, decidir los predictores del modelo basados en la evidencia disponible, sin emplear estrategias de selección de variables. El comando – `bsvalidation` – es útil tanto si el modelo es preespecificado por el investigador como si el modelo ha sido desarrollado utilizando estrategias de selección de variables automatizadas. Entre las estrategias tradicionales de selección de variables las más ampliamente utilizadas son las variantes de la selección paso a paso. Estos métodos automatizados seleccionan los predictores más importantes basados en la significación estadística. El método hacia atrás (en inglés *backward selection*) parte desde un modelo con todos los candidatos predictores, conocido como el modelo máximo, y va eliminando los predictores menos significativos; el método hacia adelante (en inglés *forward selection*) comienza desde un modelo vacío, conocido como modelo nulo, y va añadiendo los predictores más significativos (20). El comando – `bsvalidation` – incorpora las opciones `pr()` y `pe()` para fijar los umbrales de significación de exclusión (en *backward selection*) o inclusión (en *forward selection*) de una variable en el modelo . Además, ambas opciones pueden ser combinadas para establecer una estrategia de selección de variables iterativa que intercala la inclusión y eliminación de variables del modelo (97). Si se utilizan estrategias de selección de variables para construir el modelo, el método preferible es *backward selection* (19,20).

Cuando se realiza alguna estrategia de selección de variables, esta debe ser reproducida en cada muestra bootstrap. De modo que los predictores seleccionados pueden variar entre muestras. El comando – `bsvalidation` – incorpora una opción (`models`) que comunica la estructura de los modelos bootstrap finales obtenidos en cada muestra bootstrap, siguiendo la estrategia de selección de predictores especificada. Además, y por defecto, cuando se aplican estrategias de selección de variables el comando reporta el número y el porcentaje de veces que cada predictor ha sido seleccionado en el modelo final entre las *b* muestras bootstrap ejecutadas. Estos resultados permiten al investigador valorar la confianza en los predictores incluidos y excluidos del modelo pronóstico.

Los gráficos de calibración son una herramienta intuitiva que permiten al investigador interpretar los resultados del rendimiento predictivo del modelo. Un gráfico de calibración de un modelo de regresión logística presenta en el eje X las predicciones del modelo y en el eje Y el resultado de interés, que en regresión logística binaria podrá tomar valor 0 o 1. Por tal motivo las técnicas de suavizado son útiles para presentar las probabilidades observadas en relación con las probabilidades predichas por el modelo (Figura 10)(34).

La figura 10 muestra la visualización de las deficiencias en la calibración.

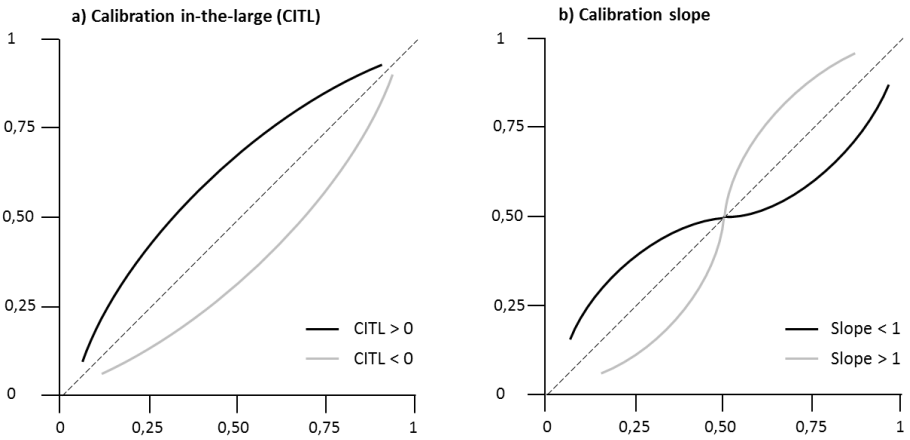


FIGURA 10. REPRESENTACIÓN GRÁFICA DE LA CALIBRACIÓN EN UN MODELO PRONÓSTICO

La recta discontinua diagonal (ángulo de 45º) representa una predicción perfecta y sirve de referencia para la interpretación de la gráfica.

En la figura 10.a) se presentan dos situaciones para la *calibration in-the-large*. La línea negra indica que el modelo sistemáticamente infraestima el riesgo del evento, es decir, produce predicciones demasiado bajas (CITL > 0). La línea gris indica que el modelo sistemáticamente sobreestima el riesgo, es decir, produce predicciones demasiado altas (CITL < 0). En la figura 10.b) se presentan dos situaciones de acuerdo a la pendiente de calibración (*calibration slope*). La línea negra indica que el modelo está sobreajustado produciendo predicciones demasiado extremas, es decir, para probabilidades altas estima predicciones demasiado altas y para probabilidades bajas estima predicciones demasiado bajas (*Slope* < 1). La línea gris indica que el modelo está infraajustado produciendo predicciones que no varían demasiado, es decir, produce predicciones que no son suficientemente altas ni bajas, demasiado próximas al promedio (*Slope* > 1).

El comando – `bsvalidation` – permite obtener un gráfico de calibración con la opción `graph`, que presenta el ajuste de las predicciones con una función de suavizado y mediante grupos de riesgo que pueden ser definidos por el usuario.

4.3.3. ARTÍCULO

Los resultados de este estudio han sido enviados con el título "*Bootstrap internal validation command for predictive logistic regression models*" a la revista Stata Journal perteneciente al primer cuartil de la categoría "Statistics & Probability".

The Stata Journal (2021)
21, Number 1, pp. 1–12 DOI: !!

Bootstrap internal validation command for predictive logistic regression models

B. M. Fernandez-Felix
Clinical Biostatistics Unit Hospital Ramón y Cajal (IRYCIS)
CIBER Epidemiology and Public Health (CIBERESP)
Madrid, Spain
borjam.fernandez@hrc.es

E. García-Esquinas
Department of Preventive
Medicine and Public Health
Autonomous University of Madrid and Idipaz
CIBERESP
Madrid, Spain

A. Muriel
Clinical Biostatistics Unit
Hospital Ramón y Cajal (IRYCIS)
CIBERESP
Madrid, Spain

A. Royuela
Biostatistics Unit
Puerta de Hierro Biomedical Research Institute
CIBERESP
Madrid, Spain

J. Zamora
Clinical Biostatistics Unit
Hospital Ramón y Cajal (IRYCIS)
CIBERESP Madrid, Spain
Institute of Metabolism and Systems Research
University of Birmingham
Birmingham, UK

Abstract. Overfitting is a common problem in the development of predictive models. It leads to an optimistic estimation of apparent model performance. Internal validation using bootstrapping techniques allows one to quantify the optimism of a predictive model and provide a more realistic estimate of its performance measures. Our objective is to build an easy-to-use command, `bsvalidation`, aimed to perform a bootstrap internal validation of a logistic regression model.

Keywords: st00!!, bsvalidation, bootstrap, internal validation, predictive model, performance, logistic, logit

1 Introduction

A multivariable predictive model is a mathematical equation that relates multiple predictors for a particular individual to the probability of future occurrence of an outcome

(Royston et al. 2009). Overfitting is a common problem in the development of these models, and it usually yields an overly optimistic model performance (Steyerberg 2009). In this context, internal validation is essential to provide a more realistic estimate of model ability to predict the risk of the outcome in a new subject. Several solutions have been proposed to correct for this optimism (sample splitting, cross-validation, and its variants leave-one-out cross-validation or leave-pair-out cross-validation). Among these strategies, bootstrapping emerges as a popular strategy to correct for optimistic estimates of the apparent performance.

The transparent reporting of a multivariable prediction model for an individual prognosis or diagnosis (TRIPOD) statement is an evidence-based guide of recommendations to standardize reporting of predictive models. The TRIPOD statement recommends bootstrapping techniques to carry out internal model validation and shrinkage methods to adjust overfitted models (Moons et al. 2015; Collins et al. 2015).

Our objective is to develop a new command, `bsvalidation`, to perform internal model validation using bootstrapping techniques that is executable as a postestimation command after the `logistic` or `logit` command. Stata has implemented postestimation commands to assess the apparent performance of the model. First, it has implemented the `lroc` postestimation command to assess model discrimination. It also has implemented `estat gof` to assess model calibration with a Hosmer–Lemeshow test. To the best of our knowledge, there is no user-defined internal validation command implemented in Stata to date such as the one we are presenting.

2 Methods

`bsvalidation` needs to be executed after `logistic` or `logit`. The command allows one to estimate different performance measures in terms of overall model fit performance (that is, how close our predictions are to the actual outcome, related to the amount of variability that is explained); discrimination (that is, how well the model distinguishes between those with and without the outcome); and calibration (that is, how well predictions and observations agree). These measures can be observed in table 1.

Table 1. Performance measures

Item	Measure	Characteristics
Overall performance (Steyerberg et al. 2010)	$\text{Brier}_{\text{scaled}}$	Range: [0, 100] High values indicate predictions are closer to the actual outcome.
Discrimination (Riley et al. 2019)	C-statistic	Range: [0.5, 1] High values indicate better discrimination.
Calibration (Riley et al. 2019)	E:O ratio	Ideal value: 1 E:O $<$ 1 indicates the model underestimates for the total number of events. E:O $>$ 1 indicates the model overestimates for the total number of events.
	Calibration-in-the-large (CITL)	Ideal value: 0 CITL $<$ 0 indicates the predictions are systematically too high. CITL $>$ 0 indicates the predictions are systematically too low.
	Calibration slope	Ideal value: 1 Slope $<$ 1 indicates the predictions are too extreme and the model is overfit. Slope $>$ 1 indicates the predictions are not varied enough and the model is underfit.

NOTE: $\text{Brier}_{\text{scaled}} = 1 - \text{Brier}_{\text{score}} / \text{Brier}_{\text{max}}$

After the user has fit a logistic predictive model in the original sample using either the `logit` or `logistic` command, the validation command goes over the following algorithm:

1. It determines its apparent performance in the original sample (table 1).
2. It draws a bootstrap sample with replacement from the original sample.
3. It builds a new prediction model–bootstrap model replicating the same modeling strategy used in the model that is being validated, and it determines its apparent performance in the bootstrap sample–bootstrap performance. If the original model is prespecified (that is, fit without variable selection), `bsvalidation` uses original model specification without any strategy for variable selection.
4. It applies the bootstrap model to the original sample to determine its performance–test performance.
5. It calculates the model's optimism as the difference between the bootstrap performance and the test performance.
6. It repeats steps 2–5 a user-defined number of times to obtain a stable averaged estimate of the optimism.
7. Finally, it subtracts the averaged optimism estimate obtained in step 6 from the initial apparent performance estimated in step 1 to obtain the optimism-corrected performance estimate.

Also, uniform shrinkage parameters—heuristic (Van Houwelingen and Le Cessie 1990) and bootstrap (Harrell 2015)—are estimated, and the coefficient of the model can be shrunk.

Our `bsvalidation` command also generates a calibration plot. Calibration is assessed using a lowess smoother function of predicted and observed risks for the overall sample. It also presents pairs of predicted and observed risks for groups defined by the user according to quantiles of predicted risk.

3 The bsvalidation command

3.1 Syntax

The syntax for `bsvalidation` is

`bsvalidation` [*varlist*] [`,` *options*]

If the final model was prespecified, *varlist* will be empty. If the model was built using selection methods (backward, forward, or stepwise), those predictors previously assessed but excluded from the final model during the selection process should be included in *varlist*.

3.2 Options

`reps(#)` specifies the number of bootstrap samples. The default is 50 samples. If you are using Stata/IC, up to 800 bootstrap samples are supported. See `help limits`.

`rseed(#)` sets the random-number seed. This option can be used to obtain reproducible results. `rseed(#)` is equivalent to typing `set seed #` prior to calling `bsvalidation`.

`adjust(`*string*`)` displays the final model after applying a uniform shrinkage factor to the regression coefficients. *string* is one of the following:

`heuristic`—uniform heuristic shrinkage parameter from Van Houwelingen and Le Cessie (1990).

`bootstrap`—uniform bootstrap shrinkage parameter from Steyerberg (2009).

`pr(#)` and `pe(#)` specify the significance level threshold for variables to be removed from or entered into the model, respectively.

`pr(#)` is backward elimination. Variables with p-value $\geq$ `pr()` are eligible to be removed.

`pe(#)` is forward selection. Variables with p-value $<$ `pe()` are eligible to be entered.

`pr(#)` and `pe(#)` indicate backward stepwise.

When a predictor-selection approach is considered, a backward elimination strategy is generally preferred (Harrell 2015).

Furthermore, `bsvalidation` displays the times each variable is selected in the final model after applying the same selection strategy for each bootstrap sample. Other variable-selection strategies such as lasso (least absolute shrinkage and selection operator) are not included in `bsvalidation`. See `help lasso`.

`models` displays the final model for each bootstrap sample. If the final model is prespecified, this option does not apply.

`eform` causes the coefficient table to be displayed in exponentiated form: for each coefficient, `exp(b)` rather than `_b` is displayed. Standard errors and confidence intervals are also transformed.

`graph` produces a calibration plot of observed against expected probabilities. Calibration is plotted in groups across the risk spectrum. Confidence intervals for the groupings are displayed as well as a lowess smoother.

This allows one to assess the calibration at the individual level. If `adjust()` is considered, then the calibration plot will be adjusted.

Other user commands to generate calibration plots can be consulted (Ensor, Snell, and Martin 2018).

`group(#)` specifies the number of percentiles to divide the predicted risks into. The default is to divide the predicted risks into 10 equally sized groups.

`min(#)` allows one to fix a lower bound of observed and expected probabilities to be plotted.

If `min()` is higher than the minimum probability predicted by the model, it is automatically rounded to the nearest first decimal to minimum.

`max(#)` allows one to fix an upper bound of observed and expected probabilities to be plotted.

If `max()` is lower than the maximum probability predicted by the model, it is automatically rounded to the nearest first decimal to maximum.

3.3 Stored results

`bsvalidation` stores the following in `e()`:

Scalars

`e(N)`	number of observations
`e(k)`	number of parameters in the final model
`e(df_m)`	degrees of freedom
`e(k_max)`	number of parameters in the maximum model
`e(boot)`	number of bootstrap samples
`e(brier)`	Brier score for model overall performance
`e(opt_brier)`	optimism of the Brier score
`e(cstat)`	C-statistic for model discrimination
`e(opt_cstat)`	optimism of the C-statistic
`e(eo_ratio)`	ratio between expected and observed events for model calibration
`e(citl)`	calibration-in-the-large for model calibration
`e(slope)`	calibration slope for model calibration
`e(heur_shrink)`	uniform heuristic shrinkage
`e(boot_shrink)`	uniform bootstrap shrinkage

Macros

`e(cmd)`	bsvalidation
`e(depvar)`	dependent variable
`e(all_vars)`	independent variables in the maximum model
`e(sel_vars)`	independent variables in the final model
`e(model)`	regression model
`e(properties)`	`b V`

Matrices

`e(b)`	coefficient vector
`e(V)`	variance–covariance matrix of the estimators

Functions

`e(sample)`	marks estimation sample

4 Examples

We illustrate the use of `bsvalidation` with a predictive model developed to estimate the risk of low birthweight using the dataset `lbw.dta` from Hosmer, Lemeshow, and Sturdivant (2013).

In the first example, the command `bsvalidation` runs a bootstrap internal validation of a prespecified model.

```
. use http://www.stata-press.com/data/r16/lbw.dta
(Hosmer & Lemeshow data)
. logistic low age lwt i.race smoke ptl ht ui
Logistic regression                             Number of obs     =        189
                                                LR chi2(8)        =      33.22
                                                Prob > chi2       =     0.0001
Log likelihood =   -100.724                     Pseudo R2         =     0.1416

------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9732636   .0354759    -0.74   0.457     .9061578    1.045339
         lwt |   .9849634   .0068217    -2.19   0.029     .9716834    .9984249
             |
        race |
       black |   3.534767   1.860737     2.40   0.016     1.259736    9.918406
       other |   2.368079   1.039949     1.96   0.050     1.001356    5.600207
             |
       smoke |   2.517698    1.00916     2.30   0.021     1.147676    5.523162
         ptl |   1.719161   .5952579     1.56   0.118     .8721455    3.388787
          ht |   6.249602   4.322408     2.65   0.008     1.611152    24.24199
          ui |     2.1351   .9808153     1.65   0.099     .8677528      5.2534
       _cons |   1.586014   1.910496     0.38   0.702     .1496092    16.8134
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
. bsvalidation, rseed(123) graph
Bootstrap sampling
..................................................   50

Apparent performance
------------------------------------------------------------------------------
                                          [95% Conf. Interval]
Overall:
        Brier scaled (%) =  16.4
Discrimination:
             C-Statistic =  0.746       0.673   0.820
Calibration:
               E:O ratio =  1.000
                    CITL = -0.000      -0.338   0.338
                   Slope =  1.000       0.613   1.387
------------------------------------------------------------------------------

Bootstrap performance (Optimism adjusted)
Number of replications:  50
------------------------------------------------------------------------------
                                       [Bootstrap 95% CI]
Overall:
        Brier scaled (%) =   5.4
Discrimination:
             C-Statistic =  0.694       0.636   0.761
Calibration:
               E:O ratio =  1.003       0.826   1.223
                    CITL =  0.000      -0.460   0.368
                   Slope =  0.712       0.455   1.037
------------------------------------------------------------------------------
```

```
Shrinkage factors
------------------------------------------------------------------------------

     Heuristic Shrinkage =  0.759
     Bootstrap shrinkage =  0.712
------------------------------------------------------------------------------
```

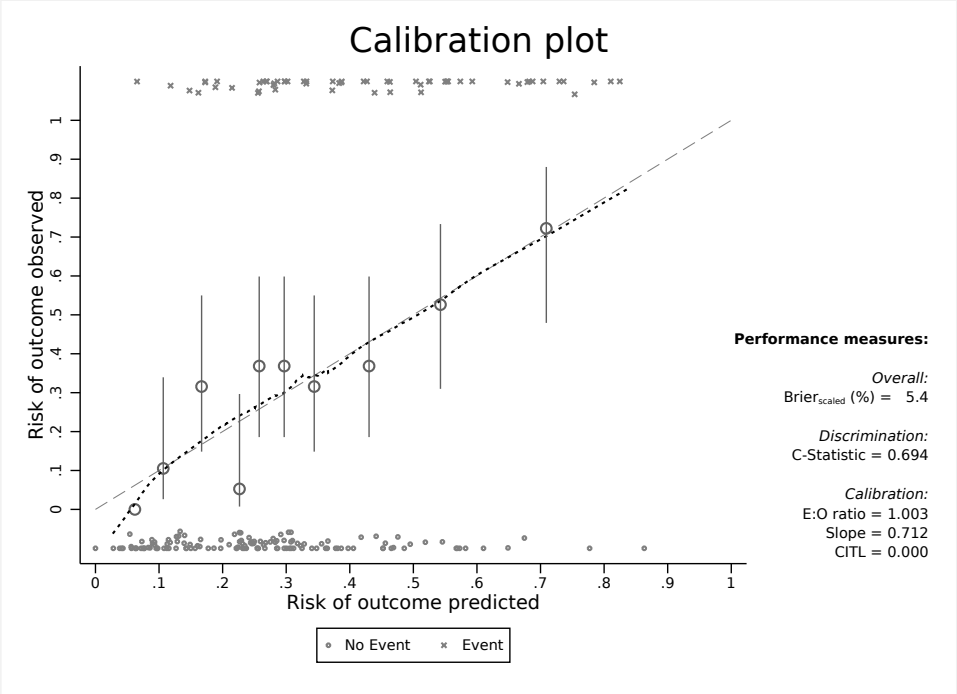


Figure 1. Calibration plot

In this first example, we fit a prespecified logistic model to predict the risk of low birthweight (defined as birthweight lower than 2,500 grams), using the mother's age (`age`), weight at last menstrual period (`lwt`), race (`race`), smoking status during pregnancy (`smoke`), previous history of premature labor (`ptl`), hypertension (`ht`), and uterine irritability (`ui`) as predictors. The `bsvalidation` output shows all apparent performance statistics (for example, C-statistic = 0.746). These performance measures are then adjusted for the estimated optimism, which is calculated from 50 (the default number) bootstrap samples (for example, C-statistic = 0.694). Additionally, by using the `graph` option, we visualize a calibration plot of observed against expected risks of low birthweight in groups defined by deciles of predicted risk, along with a smooth fitted line. Further, it shows scatterplots with the distribution of events (x symbol) and nonevents (hollow circle symbol) along the x axis.

In the second example, `bsvalidation` performs a bootstrap internal validation of a model that was previously built using a backward-selection strategy with significance level ($p = 0.1$). After the backward-selection strategy, the predictors `age` and `ptl` were

dropped. The model coefficients are finally adjusted by the bootstrap-estimated uniform shrinkage factor or coefficient.

```
. use http://www.stata-press.com/data/r16/lbw.dta, clear
(Hosmer & Lemeshow data)
. logistic low lwt i.race smoke ht ui
Logistic regression                             Number of obs     =        189
                                                LR chi2(6)        =      30.43
                                                Prob > chi2       =     0.0000
Log likelihood = -102.11978                     Pseudo R2         =     0.1297

------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lwt |   .9834361   .0066887    -2.46   0.014     .9704134    .9966336
             |
        race |
       black |   3.758631   1.959795     2.54   0.011     1.352705    10.44375
       other |   2.526023   1.087054     2.15   0.031      1.08675    5.871446
             |
       smoke |   2.817403   1.105908     2.64   0.008     1.305356    6.080917
          ht |   6.490237   4.483259     2.71   0.007     1.676009    25.13302
          ui |   2.471801   1.106213     2.02   0.043     1.028189    5.942297
       _cons |   1.054066   .9884219     0.06   0.955     .1677556    6.623063
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
. bsvalidation age ptl, rseed(123) reps(100) pr(0.1) adjust(bootstrap) eform
Bootstrap sampling
..................................................    50
..................................................   100

Apparent performance
------------------------------------------------------------------------------
                                    [95% Conf. Interval]
Overall:
        Brier scaled (%) =  15.1
Discrimination:
             C-Statistic =  0.735       0.660   0.810
Calibration:
               E:O ratio =  1.000
                    CITL = -0.000      -0.335   0.335
                   Slope =  1.000       0.600   1.400
------------------------------------------------------------------------------

Bootstrap performance (Optimism adjusted)
Number of replications: 100
------------------------------------------------------------------------------
                                    [Bootstrap 95% CI]
Overall:
        Brier scaled (%) =   4.8
Discrimination:
             C-Statistic =  0.682       0.626   0.743
Calibration:
               E:O ratio =  0.998       0.785   1.207
                    CITL =  0.009      -0.350   0.433
                   Slope =  0.712       0.484   1.039
------------------------------------------------------------------------------
```

```
Shrinkage factors
------------------------------------------------------------------------------
     Heuristic Shrinkage =  0.759
     Bootstrap shrinkage =  0.712
------------------------------------------------------------------------------

Number of times each variable is selected
------------------------------------------------------------------------------
                      Freq      %
   lwt:                 75     75.0%
   1b.race:              0      0.0%
   2.race:              87     87.0%
   3.race:              87     87.0%
   smoke:               72     72.0%
   ht:                  94     94.0%
   ui:                  62     62.0%
   age:                 21     21.0%
   ptl:                 49     49.0%
------------------------------------------------------------------------------

Model adjusted by bootstrap shrinkage
------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
low          |
         lwt |   .9881782   .0047853    -2.46   0.014     .9788434    .9976019
             |
        race |
       black |   2.566972   .9529764     2.54   0.011     1.239985    5.314051
       other |   1.934351   .5926921     2.15   0.031     1.061022    3.526521
             |
       smoke |   2.090704    .584309     2.64   0.008     1.208924    3.615647
          ht |   3.787298   1.862699     2.71   0.007     1.444391    9.930572
          ui |   1.904696    .606919     2.02   0.043      1.01999    3.556768
       _cons |   .8493539   .1397411    -0.99   0.321     .6152387    1.172556
------------------------------------------------------------------------------
```

In the second example, the model is built using a backward-selection strategy in the original data. The predictors selected in the process are `lwt`, `race`, `smoke`, `ht`, and `ui` (`logistic` command). Other candidate predictors (`age` and `ptl`) initially assessed, but excluded during the selection process, are added in the *varlist* of the `bsvalidation` command to replicate the same modeling strategy used during the development of the original model. The output shows both apparent and optimism-adjusted performance measures. Additionally, because the backward-selection strategy is replicated in each bootstrap sample, the output also shows the number of times each predictor is selected in the final model (that is, `lwt` was included in 75 out of 100 bootstrap models). Finally, the coefficients of the final model are adjusted by bootstrap-based uniform shrinkage to correct overfitting. Thus, coefficients are multiplied by 0.712.

5 Conclusion

`bsvalidation` is a useful command to run bootstrap internal validation of predictive logistic regression models. It makes this internal validation method more accessible to researchers promoting a more complete and better report of predictive models according to TRIPOD guidelines.

6 Limitations

Although `bsvalidation` helps standardize the internal validation process, a disadvantage of bootstrap validation is that it allows validation only of models built following fixed or automated modeling strategies (that is, without dynamic modeling strategies or stepwise modeling strategies). Other important steps during the modeling process, such as collapsing factor variables, assessing nonlinearities, or testing for interaction terms, cannot be handled by `bsvalidation`. The command does not handle other shrinkage methods, such as the least absolute shrinkage and selection operator (Tibshirani 1996), and cannot handle missing values.

7 Future works

In the future, we will work to solve some of the previously mentioned limitations, and we will evolve the command to validate other regression models commonly used in biomedical research, such as Cox regression.

8 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 21-2
. net install st00!!     (to install program files, if available)
. net get st00!!         (to install ancillary files, if available)
```


9 References

Collins, G. S., J. B. Reitsma, D. G. Altman, and K. G. M. Moons. 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *British Medical Journal* 350: g7594. https://doi.org/10.1136/bmj.g7594.

Ensor, J., K. I. E. Snell, and E. C. Martin. 2018. pmcalplot: Stata module to produce calibration plot of prediction model performance. Statistical Software Components S458486, Department of Economics, Boston College. https://EconPapers.repec.org/RePEc:boc:bocode:s458486.

Harrell, F. E., Jr. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. Cham, Switzerland: Springer.

Hosmer, D. W., Jr., S. Lemeshow, and R. X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley.

Moons, K. G. M., D. G. Altman, J. B. Reitsma, J. P. A. Ioannidis, P. Macaskill, E. W. Steyerberg, A. J. Vickers, D. F. Ransohoff, and G. S. Collins. 2015. Transparent

reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine* 162: W1–W73. https://doi.org/10.7326/M14-0698.

Riley, R. D. A., D. van der Windt, P. Croft, and K. G. M. Moons. 2019. *Prognosis Research in Health Care: Concepts, Methods, and Impact*. New York: Oxford University Press. https://doi.org/10.1093/med/9780198796619.001.0001.

Royston, P., K. G. M. Moons, D. G. Altman, and Y. Vergouwe. 2009. Prognosis and prognostic research: Developing a prognostic model. *British Medical Journal* 338: b604. https://doi.org/10.1136/bmj.b604.

Steyerberg, E. W. 2009. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Cham, Switzerland: Springer.

Steyerberg, E. W., A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. 2010. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* 21: 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58: 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Van Houwelingen, J. C., and S. Le Cessie. 1990. Predictive value of statistical models. *Statistics in Medicine* 9: 1303–1325. https://doi.org/10.1002/sim.4780091109.

About the authors

Borja M. Fernandez-Felix is a PhD student in the Department of Epidemiology and Public Health at the Autonomous University of Madrid. He works as a biostatistician at the Clinical Biostatistics Unit, Ramón y Cajal University Hospital, Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), in Madrid, Spain.

Esther García Esquinas works at the Department of Preventive Medicine and Public Health of the Autonomous University of Madrid.

Alfonso Muriel works as a biostatistician at the Clinical Biostatistics Unit, Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS) in Madrid, Spain.

Ana Royuela is the head of the Biostatistics Unit at Puerta de Hierro Hospital in Majadahonda, Madrid, Spain.

Javier Zamora is the head of the Clinical Biostatistics Unit at Ramón y Cajal University Hospital, and he works as a professor of biostatistics in maternal and perinatal health at the University of Birmingham, UK.

All coauthors are members of the CIBER of Epidemiology and Public Health (CIBERESP).

Discusión

5. DISCUSIÓN

En esta Tesis se presentan dos modelos pronósticos para predecir desenlaces importantes en dos contextos clínicos distintos. Ambos modelos han sido desarrollados siguiendo alternativas metodológicas complementarias. En el primero de ellos, ante la ausencia en la literatura científica de herramientas que permitieran predecir el riesgo de crisis epilépticas durante el embarazo en mujeres en tratamiento antiepiléptico, se desarrolló un modelo pronóstico *de novo* a partir de los datos primarios recogidos en el estudio EMPiRE (87). En el segundo contexto clínico, la situación es distinta y se disponía ya de varios modelos en la literatura para predecir el riesgo de mortalidad postoperatoria en pacientes con endocarditis infecciosa. Por ello se optó por hacer una revisión sistemática y meta-análisis para identificar y evaluar la calidad de los modelos disponibles y, a partir de ellos, crear un modelo único (meta-modelo) que fue optimizado para el registro nacional GAMES de endocarditis infecciosa (90). En ambos escenarios clínicos, se ha pretendido adicionalmente promover la utilización de estos modelos en la práctica clínica. Para este fin se han creado dos calculadoras online de libre acceso que han sido implementadas en la plataforma web Evidencio. Esta implementación facilita enormemente la obtención de una predicción personalizada del riesgo de los desenlaces considerados dadas las características individuales de los pacientes.

A pesar de que los métodos para el desarrollo y validación de los modelos pronóstico han sido descritos por múltiples autores (9,10,19,20,28,65), estos métodos son en ocasiones complejos para investigadores con conocimientos limitados de estadística. Por ello, es recomendable que el equipo investigador de un estudio de desarrollo o validación de un modelo pronóstico cuente entre sus miembros con algún estadístico o persona con amplio bagaje metodológico (98). Sin embargo, no siempre es así, por lo que facilitar a los investigadores herramientas útiles y de manejo sencillo, como el comando – `bsvalidation` – para el software Stata que se ha presentado en el tercer estudio de la tesis, puede ayudar a paliar los efectos de las carencias metodológicas del equipo investigador.

A continuación, y de modo individualizado se discutirá los siguientes apartados de cada aportación de la tesis: resumen de los hallazgos, comparación con la evidencia disponible, fortalezas y debilidades, implicaciones para la práctica clínica e implicaciones para la investigación. Para el tercer estudio, dado su carácter metodológico, no se discutirá el apartado de implicaciones para la práctica clínica.

5.1. ESTUDIO 1: PREDICTING SEIZURES IN PREGNANT WOMEN WITH EPILEPSY: DEVELOPMENT AND EXTERNAL VALIDATION OF A PROGNOSTIC MODEL

5.1.1. RESUMEN DE LOS HALLAZGOS

El modelo EMPiRE presenta un buen rendimiento para predecir el riesgo de convulsiones epilépticas en el momento de la visita prenatal en mujeres en tratamiento antiepiléptico. El modelo incorpora características disponibles en la práctica diaria y sencillas de obtener, tales como la edad de la mujer en la primera crisis epiléptica, el tipo de crisis, la presencia de crisis en los tres meses previos al embarazo, el estatus de salud mental, ingresos hospitalarios por crisis en embarazos previos, y la dosis de medicación antiepiléptica en el momento de la visita prenatal. El modelo mostró ser clínicamente útil en un amplio rango de probabilidades, y relevante para el seguimiento del embarazo en todos los niveles asistenciales (primaria y especializada) para identificar mujeres de alto riesgo.

5.1.2. COMPARACIÓN CON LA EVIDENCIA EXISTENTE

Hasta nuestro conocimiento se trata del primer modelo pronóstico para predecir crisis epilépticas en mujeres embarazadas. Existen dos modelos predictivos de crisis epilépticas, pero ninguno de ellos en mujeres embarazadas: Uno para predecir convulsiones en niños y adultos quienes recientemente han abandonado la medicación antiepiléptica, y otro para predecir convulsiones posteriores a una primera crisis epiléptica en individuos sin una clara indicación para comenzar tratamiento antiepiléptico (99,100).

5.1.3. FORTALEZAS Y DEBILIDADES

En el desarrollo del modelo se emplearon datos de alta calidad obtenidos en el contexto de un ensayo clínico aleatorizado multicéntrico (87). Los predictores analizados son clínicamente relevantes y accesibles en una revisión rutinaria, por lo que el modelo puede ser aplicado en la práctica clínica. El modelo no solo predice el riesgo de convulsiones durante el embarazo sino hasta 6 semanas después del parto, un periodo en el cual se incrementan los riesgos de madres y bebés (101).

El potencial de transportabilidad del modelo es elevado ya que su rendimiento fue similar en un contexto donde el control de la epilepsia en el embarazo fue distinto al seguimiento realizado en las mujeres incluidas en la cohorte de desarrollo. El modelo fue desarrollado en mujeres que habían recibido una atención sanitaria acorde a las directrices del National Institute for Health

and Care Excellence del Reino Unido, del Royal College of Obstetricians and Gynecologists y de la Scottish Intercollegiate Guidelines Network, que recomiendan el ajuste de la dosis de los fármacos antiepilépticos en base a una monitorización clínica, evitando la determinación rutinaria de los niveles de fármaco en sangre (79,102,103). Una vez desarrollado, el modelo fue externamente validado en una cohorte de mujeres que se sometieron a una monitorización rutinaria de los niveles de fármacos en sangre, con el aumento de la dosis si el nivel del fármaco descendía, acorde a las recomendaciones de la American Academy of Neurology (104).

Existen algunas limitaciones en este estudio. Las mujeres que se incorporaron al estudio lo hicieron siguiendo unos específicos criterios de inclusión que podrían limitar el uso del modelo a todas las mujeres (1,48). Siguiendo las recomendaciones en contra del uso del valproato de sodio durante el embarazo (105), en el estudio no se incorporaron mujeres con esta medicación lo que pudiera limitar la aplicabilidad en ámbitos donde esta medicación se utilice. El modelo solo se puede utilizar en mujeres tratadas con fenitoína, lamotrigina, levetiracetam o carbamazepina y cuando se dispone de la información de todos los demás predictores. Esto podría limitar su transportabilidad a países de bajos y medios ingresos con limitaciones de recursos y falta de disponibilidad de estos medicamentos (106). Aunque los resultados de la modelización fueron válidos, debe interpretarse con cautela debido al escaso número de eventos en la muestra de validación

En el modelo solo se incluyeron predictores que estuviesen disponibles en el momento de la visita rutinaria prenatal y, por tanto, no se evaluaron otras pruebas u otros factores de riesgo que podrían estar disponibles en la práctica. Estas variables adicionales podrían mejorar el rendimiento predictivo del modelo pero a costa de reducir su aplicabilidad.

5.1.4. IMPLICACIONES PARA LA PRÁCTICA CLÍNICA

Actualmente, las mujeres embarazadas con epilepsia son atendidas por varios profesionales sanitarios como médicos de atención primaria, obstetras o neurólogos. No existe una vía clínica definida para esta atención multidisciplinar. Un primer paso para lograr una atención integral es la evaluación personalizada de los riesgos de convulsiones de las gestantes. Este enfoque basado en la predicción del riesgo mediante un modelo predictivo puede ayudar a evitar muertes maternas como las que se comunican en el informe MBRRACE-UK (72).

Nos hemos abstenido de recomendar umbrales de riesgo para la toma de decisiones pues estos umbrales pueden variar en función de las decisiones y sus consecuencias (beneficios, riesgos y efectos adversos) y los costos de la intervención. Por ejemplo, un médico de atención primaria

podría considerar un riesgo del 20% como un umbral adecuado para limitar a la gestante la conducción de un vehículo a motor o para derivar a ésta de forma temprana a una atención especializada neurológica. Si la capacidad de conducir es crucial para la madre para acudir a su trabajo y/u otras responsabilidades, después de hablar con los médicos, se podría optar por elegir un umbral más alto del 20%. Sin embargo, los médicos de atención especializada podrían requerir umbrales todavía más altos cuando la decisión que se debe tomar implica un control antenatal frecuente (semanal o quincenal), o el uso durante el parto de intervenciones invasivas para aliviar el dolor, como la epidural, o un control posnatal más exhaustivo. En cualquier caso, nuestro análisis de la curva de decisión muestra que el modelo es útil para la toma de decisiones en una amplia gama de umbrales de probabilidad. El conocimiento del riesgo de convulsiones además puede minimizar la falta de adherencia a la medicación durante el embarazo, uno de los principales factores que explican el incremento de las convulsiones durante el embarazo (74,75,107).

El uso del modelo EMPiRE ha sido recientemente recomendado en la guía del Royal College of Obstetrics and Gynecology para los servicios de medicina materna en la pandemia del coronavirus (COVID-19) (108).

5.1.5. IMPLICACIONES PARA LA INVESTIGACIÓN

El modelo EMPiRE se ha desarrollado siguiendo la metodología recomendada tanto para su desarrollo y validación (19,20,29) como para su comunicación y presentación del modelo (9,10). Los valores perdidos de los predictores se trataron mediante imputación múltiple, evitando así la pérdida de información útil (88,89). Ajustamos el potencial exceso de optimismo y abordamos los problemas relacionados con el sobreajuste en el modelo, combinando de forma eficiente la metodología de selección de predictores y ajuste de sus coeficientes con las técnicas de remuestreo bootstrapping y los métodos de imputación múltiple. La combinación de estas estrategias metodológicas es un campo en estudio y que, actualmente, supone una alta carga computacional en el desarrollo y validación de los modelos. El modelo ha sido presentado como como un nomograma y una calculadora online que permiten su uso inmediato para predecir el riesgo de convulsiones sin la necesidad de hacer cálculos. Proporcionamos información sobre su uso clínico para varios umbrales de probabilidad en la toma de decisiones.

La validación externa de este modelo debería extenderse a diferentes contextos sanitarios y poblaciones para evaluar plenamente la transportabilidad del modelo (16). Debe evaluarse también el impacto del uso del modelo EMPiRE en la práctica clínica mediante ensayos clínicos

aleatorizados, quizás en aleatorización en clusters, para evaluar si el empleo del modelo ayuda a mejorar los resultados tales como el período libre de convulsiones o la calidad de vida de estas mujeres. Si bien se espera que la herramienta mejore el conocimiento de las mujeres sobre su estado de riesgo de convulsiones durante el embarazo, el efecto del modelo EMPiRE en los niveles de ansiedad de las mujeres no se conoce y debería también ser evaluado.

Se necesitan más estudios para valorar la aceptabilidad de la herramienta por parte de las mujeres con epilepsia y las instituciones de atención médica, sus umbrales preferidos de elección y los costes derivados de las decisiones de casos falsos positivos y falsos negativos.

5.2. ESTUDIO 2: PROGNOSTIC MODELS FOR MORTALITY AFTER CARDIAC SURGERY IN PATIENTS WITH INFECTIVE ENDOCARDITIS: A SYSTEMATIC REVIEW AND AGGREGATION OF PREDICTION MODELS

5.2.1. RESUMEN DE LOS HALLAZGOS

En la revisión sistemática de modelos de predicción para mortalidad postoperatoria en pacientes con endocarditis infecciosa identificamos y evaluamos críticamente 11 modelos que habían sido comunicados en 9 estudios. De los once modelos identificados en la búsqueda tan sólo 3 no fueron clasificados con alto riesgo de sesgo. La mayoría de los modelos con alto riesgo de sesgo presentaron carencias en los métodos estadísticos lo cual sugiere que los rendimientos predictivos comunicados son excesivamente optimistas, y en futuros sujetos rendirá peor de lo que los investigadores han informado. Los tamaños de muestra utilizados para desarrollar los modelos fueron limitados y, es bien sabido, que este es un problema que conduce a predicciones inexactas y, en consecuencia, decisiones de salud incorrectas en la práctica (109).

Solo cuatro de los 11 modelos identificados comunicaron la ecuación completa del modelo, necesaria para poder validarlos externamente. Las ecuaciones de dos modelos se recuperaron previa solicitud a los autores de correspondencia. Los tres modelos que fueron clasificados con bajo o incierto riesgo de sesgo fueron agregados para construir un meta-modelo.

Nuestro meta-modelo incluyó los siguientes predictores: edad, sexo, insuficiencia renal, cirugía cardíaca previa, enfermedad pulmonar crónica, hipertensión pulmonar, fracción de eyección del ventrículo izquierdo, estado preoperatorio crítico, clasificación del estado funcional de la New York Heart Association (NYHA), presencia de complicaciones paravalvulares (absceso y/o fístula), urgencia del procedimiento, número de válvulas/prótesis tratadas, ubicación de la válvula y la etiología de la infección. El meta-modelo mostró mejor rendimiento predictivo que los modelos originales y fue internamente validado mediante técnicas bootstrapping. Los resultados indican que no hubo un exceso de optimismo. La muestra de validación utilizada fue lo suficientemente grande como para permitir combinar y actualizar los modelos publicados. Por lo tanto, es muy probable que el meta-modelo sea menos propenso a un exceso de optimismo en el rendimiento, y estimar predicciones más generalizables a nuevas poblaciones o entornos de pacientes, ya que fue construido a partir de la evidencia de varias cohortes de pacientes y optimizado para los datos un registro nacional. El riesgo de mortalidad postoperatoria puede ser calculado fácilmente utilizando la herramienta on-line de libre acceso.

5.2.2. COMPARACIÓN CON LA EVIDENCIA EXISTENTE

Esta es la primera revisión sistemática de modelos de predicción de mortalidad postoperatoria en pacientes con endocarditis infecciosa. La mayoría de los estudios de desarrollo de nuevos modelos pronósticos se basan en tamaños de muestra pequeños y las estrategias de modelado están excesivamente dirigidas por los datos disponibles sin considerar el conocimiento previo, lo que lleva a modelos ineficientes. Otros autores han realizaron estudios de validación externa de varios modelos incluidos en esta revisión sistemática pero ninguno de ellos realizó una valoración crítica de los mismos (110–113). En un estudio anterior, Varela *et al.* desarrollaron un modelo de pronóstico (APORTEI *score*) basado en una revisión sistemática de factores pronóstico (no modelos) asociados con la mortalidad hospitalaria. Evaluaron un total de 11 potenciales factores pronósticos preoperatorios relacionados con las condiciones prequirúrgicas, todos ellos han sido incluidos en nuestro meta-modelo. El modelo APORTEI se construyó utilizando múltiples metanálisis univariados de las asociaciones sin ajustar (crudas) de cada predictor con la mortalidad, sin considerar posibles correlaciones entre covariables. El peso de cada variable incluida en el modelo se obtuvo dividiendo el coeficiente de regresión combinado de cada variable por el coeficiente de regresión combinado más pequeño, redondeado al número entero más cercano. De modo que los coeficientes de regresión se transformaron en puntos de riesgo para crear el *score* APORTEI (114,115).

5.2.3. FORTALEZAS Y DEBILIDADES

En el desarrollo del meta-modelo se ha incorporado la evidencia disponible de modelos previamente publicados. Los predictores analizados, por tanto, eran clínicamente relevantes y accesibles antes del proceso quirúrgico, por lo que el modelo puede ser aplicado en la práctica clínica antes de derivar a un paciente a cirugía.

Solo han sido agregados los modelos de predicción que fueron valorados con bajo o incierto riesgo de sesgo, descartando los modelos de baja calidad que presentaban carencias en el diseño y/o análisis. El meta-modelo resultante incorporó el conocimiento previo de manera óptima y mejoró el rendimiento predictivo de los modelos existentes.

Nuestro estudio tiene algunas limitaciones. La definición del desenlace en el conjunto de datos de validación fue la mortalidad postoperatoria (intrahospitalaria o a los 30 días), y la definición del desenlace en los tres modelos utilizados para la agregación fue la mortalidad a los 30 días. A pesar de esta diferencia, un análisis de sensibilidad mostró que el meta-modelo superó a todos

los modelos publicados cuando exploramos sus rendimientos predictivos para la mortalidad a 30 días. Dos de los tres modelos considerados para la agregación se desarrollaron en la misma cohorte de pacientes. Esta circunstancia aumenta la probabilidad de que se incluyan los mismos predictores en ambos modelos y, por lo tanto, podría magnificar su asociación con el desenlace en el meta-modelo. El meta-modelo no incluyó algunos predictores asociados con la mortalidad post-operatoria de los estudios que no comunicaron la ecuación del modelo o que habían sido clasificados con alto riesgo de sesgo. Estas variables adicionales podrían mejorar el rendimiento del meta-modelo. Aunque la definición de predictores en la muestra de validación (GAMES) estaba estandarizada, estas podrían diferir de las definiciones de los estudios primarios.

5.2.4. IMPLICACIONES PARA LA PRÁCTICA CLÍNICA

La decisión de realizar una cirugía en la endocarditis infecciosa sigue siendo un desafío en la práctica clínica y debe surgir después de un cuidadoso equilibrio entre el riesgo del procedimiento y su beneficio estimado.

El riesgo de mortalidad varía considerablemente en pacientes afectados de endocarditis infecciosa que requirieron tratamiento quirúrgico dependiendo de varios factores tales como las características clínicas del paciente, el estado preoperatorio o la estrategia de tratamiento. El estado crítico preoperatorio es un factor de mortalidad bien conocido, por lo que los pacientes con fracción de eyección del ventrículo izquierdo (FEVI) deprimida, NYHA > I, insuficiencia renal, así como pacientes que se someten a intervenciones urgentes o emergentes tienen peor pronóstico. Además, la agresividad de la infección y las dificultades técnicas quirúrgicas también implican un mayor riesgo de mortalidad. Por eso es esperable un peor pronóstico en pacientes con endocarditis infecciosa por microorganismos agresivos (como *Staphylococcus spp.* u hongos) o en pacientes con abscesos paravalvulares, fístulas o cirugía cardíacas previas, ya que el abordaje quirúrgico en estos pacientes es un desafío.

Aunque las estimaciones del riesgo predicho de la mortalidad no ayudan a establecer las cargas de la futilidad quirúrgica, aportan un gran valor al ayudar a los equipos especialista en endocarditis a controlar esta compleja enfermedad.

Las directrices para el seguimiento de la endocarditis infecciosa de 2015 (86), recomiendan el score de riesgo creado por De Feo-Cotrufo *et al.* para la endocarditis infecciosa nativa (116). Es esperable que posteriores actualizaciones de estas recomendaciones incluyan las nuevas herramientas pronósticas específicas de endocarditis infecciosa. El meta-modelo generado

debería ser tomado en consideración pues recoge la evidencia disponible y ha mostrado un mejor rendimiento predictivo que los modelos individuales.

5.2.5. IMPLICACIONES PARA LA INVESTIGACIÓN

La revisión sistemática alerta de la baja calidad en el diseño y en los análisis estadísticos de los estudios de modelos pronóstico para endocarditis infecciosa. Este déficit no es exclusivo de este campo de la medicina, sino que ya ha sido reseñado por otros investigadores en diferentes ámbitos (8,38).

Aunque el sobreajuste es un problema menor en los métodos de agregación de modelos pues la cantidad de parámetros desconocidos que son estimados suele ser menor que cuando desarrollamos un modelo desde cero (117). Es necesario realizar múltiples validaciones externas para apreciar plenamente la validez del meta-modelo (67). El impacto clínico del meta-modelo debe ser estudiado mediante técnicas de curva de decisión que permitan evaluar la utilidad del modelo a lo largo del espectro total de probabilidades basado en las consecuencias de los falsos positivos y falsos negativos (27,52). Se necesitan más estudios para valorar la aceptabilidad de la herramienta por parte de los cirujanos y los servicios de cirugía cardiaca.

Métodos de agregación que consideren la incertidumbre en los coeficientes de los predictores, así como en las estimaciones de los pesos de cada modelo en la muestra de validación deben ser explorados en el futuro. Los estudios de simulación que evalúen el impacto de la agregación de modelos en diferentes escenarios clínicos ayudarán a mejorar el desarrollo de meta-modelos basados en la evidencia de estudios existentes.

5.3. ESTUDIO 3: BOOTSTRAP INTERNAL VALIDATION COMMAND FOR PREDICTIVE LOGISTIC REGRESSION MODELS

5.3.1. RESUMEN DE LOS HALLAZGOS

En este estudio se ha desarrollado un comando de análisis para el software estadístico Stata. El comando – `bsvalidation` – permite realizar la validación interna de modelos de regresión logística binaria usando técnicas bootstrap.

El comando calcula las medidas del rendimiento global, discriminación y calibración en términos aparentes y ajustados por el exceso de optimismo. Además, permite generar gráficos de calibración usando una función de suavizado de los riesgos predichos y observados, y por grupos definidos por el usuario según los cuantiles de riesgo pronosticado. El programa calcula los parámetros uniformes de penalización que pueden ser utilizados para ajustar los coeficientes del modelo en presencia de sobreajuste.

5.3.2. COMPARACIÓN CON LA EVIDENCIA EXISTENTE

El software Stata incorpora varios comandos post-estimación que calculan medidas del rendimiento del modelo. Sin embargo, estos comandos sólo permiten estimar medidas del rendimiento aparente. Hasta nuestro conocimiento no existe ningún comando que permita realizar la validación interna mediante técnicas bootstrap y obtener el rendimiento del modelo corregido por el optimismo. Hasta la fecha el único modo para obtener los resultados que reporta el comando – `bsvalidation` – es a través de un manejo sofisticado de las técnicas de sintaxis, lo cual limita las posibilidades de los investigadores y usuarios legos en esta materia.

El software estadístico R (118) también tiene la posibilidad de instalar paquetes desarrollados por los usuarios (como `bsvalidation`), y entre los más reconocidos por los usuarios de este software destaca el paquete `rms` (119) desarrollado por Harrell Jr. y que ofrece opciones de validación interna muy similares a las presentadas en el comando – `bsvalidation` –.

5.3.3. FORTALEZAS Y DEBILIDADES

El comando – `bsvalidation` – permite estimar las medidas recomendadas para comunicar los resultados del rendimiento predictivo de un modelo pronóstico de regresión logística. Presenta medidas del rendimiento global, la calibración y la discriminación. Se trata de un comando post-estimación que se emplea de modo sencillo tras ejecutar el modelo final usando los comandos

`logistic` o `logit` de Stata. De forma simultánea se presentan los resultados de la validación aparente y de la validación bootstrap corregidos por el exceso de optimismo. Entre las opciones que incorpora el comando, se pueden ajustar diversas estrategias automatizadas de selección de variables. Así, como testar la fiabilidad de los predictores incluidos en el modelo final a través de la frecuencia con la que cada predictor fue incluido en los modelos de las muestras bootstrap. En caso, de presencia de sobreajuste el usuario dispone de opciones que permiten penalizar los coeficientes del modelo por varios factores de contracción. Finalmente, se pueden presentar los gráficos de calibración.

El comando presenta algunas limitaciones y a pesar de que – `bsvalidation` – ayuda a estandarizar el proceso de validación interna, una desventaja es que permite la validación de modelos cuyo diseño ha sido preespecificado sin seguir estrategias de selección de variables o para modelos con una estrategia de selección automatizada. Esta circunstancia hace que estrategias de selección más dinámicas tales como la agregación de factores o la evaluación de relaciones no lineales o de términos de interacción no sean consideradas en la validación del modelo. Recientes estrategias que combinan la selección de variables con la estimación penalizada de los coeficientes del modelo como LASSO (*Least Absolute Shrinkage and Selection Operator*) (45) no han sido implementadas en el comando de validación. El comando – `bsvalidation` – no funciona cuando el modelo es estimado usando múltiples conjuntos de datos de imputación en presencia de valores faltantes.

5.3.4. IMPLICACIONES PARA LA INVESTIGACIÓN

El comando – `bsvalidation` – ha sido implementado en el software Stata y puede ser descargado por los usuarios desde su propio repositorio (`net install bsvalidation`). El comando puede ser un punto de partida para asegurar un reporte completo de los modelos pronósticos de regresión logística. Dada la sencillez de su uso el comando podrá ser utilizado por usuarios con un nivel básico de programación o del manejo del software.

En futuras actualizaciones el comando deberá incorporar algunas de las limitaciones presentes en la versión actual, como las estrategias de modelización LASSO o la combinación de estas estrategias con los métodos de imputación múltiple.

Una herramienta similar a – `bsvalidation` – es necesaria para otros modelos de regresión frecuentemente empleados en investigación clínica y en salud pública, tales como el modelo de regresión de Cox que permite modelizar el tiempo hasta que se produce el desenlace de interés.

CONCLUSIONES

6. CONCLUSIONES

6.1. CONCLUSIONES CLÍNICAS

1. El uso del modelo EMPiRE promueve una mejor atención sanitaria de la enfermedad epiléptica durante el embarazo, permitiendo individualizar las decisiones clínicas de acuerdo a las características de la mujer.

2. El uso del meta-modelo para la predicción del riesgo de mortalidad en endocarditis infecciosa contribuye a mejorar la atención clínica de la enfermedad, permitiendo personalizar las decisiones de acuerdo a las características individuales del paciente.

6.2. CONCLUSIONES METODOLÓGICAS

1. Los métodos estadísticos que combinan técnicas de imputación de datos faltantes con métodos novedosos de selección y ajuste de predictores son óptimos para obtener predicciones válidas del riesgo de crisis epilépticas en mujeres embarazadas.

2. La revisión sistemática de los estudios de desarrollo de modelos pronósticos en endocarditis refleja el alto riesgo de sesgo en el diseño y/o análisis de los modelos de mortalidad post-operatoria existentes.

3. Los métodos de regresión apilada combinados con datos primarios de una muestra de validación pueden ser una alternativa eficaz para desarrollar un metamodelo basado en la agregación de los modelos existentes.

4. El comando – `bsvalidation` – es una herramienta útil para realizar la validación interna de los modelos predictivos siguiendo las técnicas de remuestreo bootstrapping que permite a los investigadores comunicar los resultados del rendimiento predictivo según los estándares recomendados.

BIBLIOGRAFÍA

7. BIBLIOGRAFÍA

1. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 23 de febrero de 2009;338(feb23 1):b375-b375.

2. Chauffard A. Address in Medicine, ON MEDICAL PROGNOSIS: ITS METHODS, ITS EVOLUTION, ITS LIMITATIONS: Delivered at the Seventeenth International Congress of Medicine. BMJ. 9 de agosto de 1913;2(2745):286-90.

3. Hutchison R. An Address on THE PRINCIPLES OF DIAGNOSIS. BMJ. 3 de marzo de 1928;1(3504):335-7.

4. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. BMJ. 24 de abril de 2013;346(feb05 1):e5595-e5595.

5. Riley RD, Hayden JA, Steyerberg EW, Moons KGM, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. PLoS Med. 5 de febrero de 2013;10(2):e1001380.

6. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Med. 5 de febrero de 2013;10(2):e1001381.

7. Hingorani AD, Windt DA v. d., Riley RD, Abrams K, Moons KGM, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: Stratified medicine research. BMJ. 24 de abril de 2013;346(feb05 1):e5793-e5793.

8. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 7 de abril de 2020;m1328.

9. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Ann Intern Med. 19 de mayo de 2015;162(10):735-6.

10. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 6 de enero de 2015;162(1):W1.

11. Laupacis A. Users' Guides to the Medical Literature: V. How to Use an Article About Prognosis. JAMA. 20 de julio de 1994;272(3):234.

12. Kappen TH, van Loon K, Kappen MAM, van Wolfswinkel L, Vergouwe Y, van Klei WA, et al. Barriers and facilitators perceived by physicians when using prediction models in practice. J Clin Epidemiol. febrero de 2016;70:136-45.

13. Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. BMJ. 31 de marzo de 2009;338:b604.

14. Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JPA, et al. Biomedical research: increasing value, reducing waste. The Lancet. enero de 2014;383(9912):101-4.

15. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 7 de febrero de 2006;144(3):201-9.

16. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. marzo de 2015;68(3):279-89.

17. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 29 de febrero de 2000;19(4):453-73.

18. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338(7708):1432-5.

19. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating [Internet]. Cham: Springer International Publishing; 2019 [citado 7 de abril de 2021]. (Statistics for Biology and Health). Disponible en: http://link.springer.com/10.1007/978-3-030-16399-0

20. Harrell , FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis [Internet]. Cham: Springer International Publishing; 2015 [citado 7 de abril de 2021]. (Springer Series in Statistics). Disponible en: http://link.springer.com/10.1007/978-3-319-19425-7

21. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. diciembre de 1996;49(12):1373-9.

22. Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol. agosto de 2016;76:175-82.

23. van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol. diciembre de 2016;16(1):163.

24. van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. Stat Methods Med Res. agosto de 2019;28(8):2455-74.

25. Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med. 30 de marzo de 2019;38(7):1276-96.

26. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. diciembre de 2014;14(1):40.

27. 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. Epidemiol Camb Mass. enero de 2010;21(1):128-38.

28. Riley RD, Windt D van der, Croft P, Moons KGM. Prognosis research in healthcare: concepts, methods, and impact [Internet]. 2019 [citado 7 de abril de 2021]. Disponible en: https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5891544

29. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 1 de agosto de 2014;35(29):1925-31.

30. Cox DR, Snell EJ. Analysis of Binary Data. 2.ª ed. London: Chapman & Hall; 1989.

31. Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika. 1991;78:691-2.

32. Brier G. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. Mon Weather Rev. 1950;78:1-3.

33. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York: Wiley; 2000. 373 p. (Wiley series in probability and statistics).

34. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med. 10 de febrero de 2014;33(3):517-35.

35. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. abril de 1982;143(1):29-36.

36. for the IPPIC Collaborative Network, Snell KIE, Allotey J, Smuk M, Hooper R, Chan C, et al. External validation of prognostic models predicting pre-eclampsia: individual participant data meta-analysis. BMC Med. diciembre de 2020;18(1):302.

37. Onland W, Debray TP, Laughon MM, Miedema M, Cools F, Askie LM, et al. Clinical prediction models for bronchopulmonary dysplasia: a systematic review and external validation study. BMC Pediatr. diciembre de 2013;13(1):207.

38. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. BMC Med Inform Decis Mak. diciembre de 2006;6(1):38.

39. Counsell C, Dennis M. Systematic Review of Prognostic Models in Patients with Acute Stroke. Cerebrovasc Dis. 2001;12(3):159-70.

40. Steyerberg EW. Validation in prediction research: the waste by data splitting. J Clin Epidemiol. noviembre de 2018;103:131-3.

41. Refaeilzadeh P, Tang L, Liu H. Cross-Validation. En: Liu L, Özsu MT, editores. Encyclopedia of Database Systems [Internet]. Boston, MA: Springer US; 2009 [citado 14 de mayo de 2021]. p. 532-8. Disponible en: http://link.springer.com/10.1007/978-0-387-39940-9_565

42. Efron B. Bootstrap methods: another look at the jackknife. Ann Stat. 1979;7:1-26.

43. Brunelli A. A synopsis of resampling techniques. J Thorac Dis. diciembre de 2014;6(12):1879-82.

44. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med. noviembre de 1990;9(11):1303-25.

45. Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267-88.

46. Ensor J, Snell KIE, Debray TPA, Lambert PC, Look MP, Mamas MA, et al. Individual participant data meta-analysis for external validation, recalibration, and updating of a flexible parametric prognostic model. Stat Med. 15 de junio de 2021;40(13):3066-84.

47. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 22 de junio de 2016;353:i3140.

48. Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart Br Card Soc. mayo de 2012;98(9):683-90.

49. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart Br Card Soc. mayo de 2012;98(9):691-8.

50. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol. enero de 2008;61(1):76-86.

51. Toll DB, Janssen KJM, Vergouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol. noviembre de 2008;61(11):1085-94.

52. Vickers AJ, Holland F. Decision curve analysis to evaluate the clinical benefit of prediction models. Spine J Off J North Am Spine Soc. 3 de marzo de 2021;

53. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 16 de mayo de 2016;i2416.

54. Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. BMJ. 28 de noviembre de 2011;343(nov28 1):d7163-d7163.

55. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions [Internet]. 2020 [citado 29 de abril de 2021]. Disponible en: https://doi.org/10.1002/9781119536604

56. Urrutia G, Bonfill X. [Systematic reviews: a key tool for clinical and health decision making]. Rev Esp Salud Publica. febrero de 2014;88(1):1-3.

57. Roqué M, Martínez-García L, Solà I, Alonso-Coello P, Bonfill X, Zamora J. Toolkit of methodological resources to conduct systematic reviews. F1000Research. 14 de octubre de 2020;9:82.

58. https://methods.cochrane.org/prognosis/.

59. https://www.crd.york.ac.uk/prospero/.

60. Geersing G-J, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons K. Search Filters for Finding Prognostic and Diagnostic Prediction Studies in Medline to Enhance Systematic Reviews. Smalheiser NR, editor. PLoS ONE. 29 de febrero de 2012;7(2):e32844.

61. Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. J Am Med Inform Assoc JAMIA. diciembre de 2002;9(6):653-8.

62. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. octubre de 2009;62(10):1006-12.

63. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ. 2 de enero de 2015;350:g7647.

64. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. PLoS Med. 14 de octubre de 2014;11(10):e1001744.

65. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med. 1 de enero de 2019;170(1):W1.

66. Riley RD, Moons KGM, Snell KIE, Ensor J, Hooft L, Altman DG, et al. A guide to systematic review and meta-analysis of prognostic factor studies. BMJ. 30 de enero de 2019;k4597.

67. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 5 de enero de 2017;i6460.

68. Debray TPA, Koffijberg H, Nieboer D, Vergouwe Y, Steyerberg EW, Moons KGM. Meta-analysis and aggregation of multiple published prediction models: Meta-analysis and aggregation of multiple published prediction models. Stat Med. 30 de junio de 2014;33(14):2341-62.

69. Debray TPA, Koffijberg H, Vergouwe Y, Moons KGM, Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. Stat Med. 15 de octubre de 2012;31(23):2697-712.

70. Edey S, Moran N, Nashef L. SUDEP and epilepsy-related mortality in pregnancy. Epilepsia. julio de 2014;55(7):e72-74.

71. Knight M, Nair M, Tuffnell D, Kenyon S, Shakespeare J, Brocklehurst P, et al., editors. Saving lives, improving mothers' care—surveillance of maternal deaths in the UK 2012–14 and lessons learned to inform maternity care from the UK and Ireland Confidential Enquiries into Maternal Deaths and Morbidity 2009–14. Oxford: University of Oxford National Perinatal Epidemiology Unit; 2016.

72. Knight M, Kenyon S, Brocklehurst P, Neilson J, Shakespeare J, Kurinczuk JJ, editors. Saving lives, improving mothers' care: lessons learned to inform future maternity care from the UK and Ireland Confidential Enquiries into Maternal Deaths and Morbidity 2009–12. Oxford: University of Oxford National Perinatal Epidemiology Unit; 2014.

73. Cantwell R, Clutton-Brock T, Cooper G, Dawson A, Drife J, Garrod D, et al. Saving Mothers' Lives: Reviewing maternal deaths to make motherhood safer: 2006-2008. The Eighth Report of the Confidential Enquiries into Maternal Deaths in the United Kingdom. BJOG Int J Obstet Gynaecol. marzo de 2011;118 Suppl 1:1-203.

74. Man S-L, Petersen I, Thompson M, Nazareth I. Antiepileptic drugs during pregnancy in primary care: a UK population based study. PloS One. 2012;7(12):e52339.

75. Nordeng H, Ystrøm E, Einarson A. Perception of risk regarding the use of medications and other exposures during pregnancy. Eur J Clin Pharmacol. febrero de 2010;66(2):207-14.

76. Charyton C, Elliott JO, Lu B, Moore JL. The impact of social support on health related quality of life in persons with epilepsy. Epilepsy Behav EB. diciembre de 2009;16(4):640-5.

77. Loring DW, Meador KJ, Lee GP. Determinants of quality of life in epilepsy. Epilepsy Behav EB. diciembre de 2004;5(6):976-80.

78. Smeets VMJ, van Lierop BAG, Vanhoutvin JPG, Aldenkamp AP, Nijhuis FJN. Epilepsy and employment: literature review. Epilepsy Behav EB. mayo de 2007;10(3):354-62.

79. Royal College of Obstetricians and Gynaecologists Epilepsy in pregnancy. Green-top Guideline No. 68. London: Royal College of Obstetricians and Gynaecologists; 2016.

80. Murdoch DR. Clinical Presentation, Etiology, and Outcome of Infective Endocarditis in the 21st Century: The International Collaboration on Endocarditis–Prospective Cohort Study. Arch Intern Med. 9 de marzo de 2009;169(5):463.

81. Thuny F, Grisoli D, Collart F, Habib G, Raoult D. Management of infective endocarditis: challenges and perspectives. The Lancet. marzo de 2012;379(9819):965-75.

82. Okada K, Okita Y. Surgical treatment for aortic periannular abscess/pseudoaneurysm caused by infective endocarditis. Gen Thorac Cardiovasc Surg. abril de 2013;61(4):175-81.

83. Spiliopoulos K, Haschemi A, Fink G, Kemkes B-M. Infective endocarditis complicated by paravalvular abscess: a surgical challenge. An 11-year single center experience. Heart Surg Forum. abril de 2010;13(2):E67-73.

84. Anguera I, del Río A, Moreno A, Paré C, Mestres CA, Miró JM. Complications of native and prosthetic valve infective endocarditis: update in 2006. Curr Infect Dis Rep. junio de 2006;8(4):280-8.

85. Horstkotte D, Follath F, Gutschik E, Lengyel M, Oto A, Pavie A, et al. Guidelines on prevention, diagnosis and treatment of infective endocarditis executive summary; the task force on infective endocarditis of the European society of cardiology. Eur Heart J. febrero de 2004;25(3):267-76.

86. Habib G, Lancellotti P, Antunes MJ, Bongiorni MG, Casalta J-P, Del Zotti F, et al. 2015 ESC Guidelines for the management of infective endocarditis: The Task Force for the Management of Infective Endocarditis of the European Society of Cardiology (ESC)Endorsed by: European Association for Cardio-Thoracic Surgery (EACTS), the European Association of Nuclear Medicine (EANM). Eur Heart J. 21 de noviembre de 2015;36(44):3075-128.

87. Thangaratinam S, Marlin N, Newton S, Weckesser A, Bagary M, Greenhill L, et al. AntiEpileptic drug Monitoring in PREgnancy (EMPiRE): a double-blind randomised trial on effectiveness and acceptability of monitoring strategies. Health Technol Assess Winch Engl. mayo de 2018;22(23):1-152.

88. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.

89. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 20 de febrero de 2011;30(4):377-99.

90. Muñoz P, Kestler M, De Alarcon A, Miro JM, Bermejo J, Rodríguez-Abella H, et al. Current Epidemiology and Outcome of Infective Endocarditis: A Multicenter, Prospective, Cohort Study. Medicine (Baltimore). octubre de 2015;94(43):e1816.

91. Fernández-Hidalgo N, Ferreria-González I, Marsal JR, Ribera A, Aznar ML, de Alarcón A, et al. A pragmatic approach for mortality prediction after surgery in infective endocarditis: optimizing and refining EuroSCORE. Clin Microbiol Infect. octubre de 2018;24(10):1102.e7-1102.e15.

92. Di Mauro M, Dato GMA, Barili F, Gelsomino S, Santè P, Corte AD, et al. A predictive model for early mortality after surgical treatment of heart valve or prosthesis infective endocarditis. The EndoSCORE. Int J Cardiol. agosto de 2017;241:97-102.

93. Breiman L. Stacked regressions. Mach Learn. julio de 1996;24(1):49-64.

94. StataCorp. 2019. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC.

95. Copas JB. Regression, Prediction and Shrinkage. J R Stat Soc Ser B Methodol. 1983;45(3):311-54.

96. Efron B, Morris C. Stein's Paradox in Statistics. Sci Am. 1977;236(5):119-27.

97. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. Biom J. mayo de 2018;60(3):431-49.

98. Li Z, Wang X, Barnhart HX, Wang Y. Working With Statisticians in Clinical Research. Stroke [Internet]. noviembre de 2018 [citado 16 de mayo de 2021];49(11). Disponible en: https://www.ahajournals.org/doi/10.1161/STROKEAHA.118.022266

99. Kim LG, Johnson TL, Marson AG, Chadwick DW, MRC MESS Study group. Prediction of risk of seizure recurrence after a single seizure and early epilepsy: further results from the MESS trial. Lancet Neurol. abril de 2006;5(4):317-22.

100. Lamberink HJ, Otte WM, Geerts AT, Pavlovic M, Ramos-Lizana J, Marson AG, et al. Individualised prediction model of seizure recurrence and long-term outcomes after withdrawal of antiepileptic drugs in seizure-free patients: a systematic review and individual participant data meta-analysis. Lancet Neurol. julio de 2017;16(7):523-31.

101. Aguglia U, Barboni G, Battino D, Cavazzuti GB, Citernesi A, Corosu R, et al. Italian consensus conference on epilepsy and pregnancy, labor and puerperium. Epilepsia. enero de 2009;50 Suppl 1:7-23.

102. National Institute for Health and Care Excellence. Epilepsies, diagnosis and management. Clinical guideline CG137. London: National Institute for Clinical Excellence; 2018.

103. Scottish Intercollegiate Guidelines Network. Diagnosis and management of epilepsy in adults. SIGN 143. Edinburgh: Scottish Intercollegiate Guidelines Network; 2018.

104. Harden CL, Hopp J, Ting TY, Pennell PB, French JA, Hauser WA, et al. Practice parameter update: management issues for women with epilepsy--focus on pregnancy (an evidence-based review): obstetrical complications and change in seizure frequency: report of the Quality Standards Subcommittee and Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology and American Epilepsy Society. Neurology. 14 de julio de 2009;73(2):126-32.

105. Bromley R, Weston J, Adab N, Greenhalgh J, Sanniti A, McKay AJ, et al. Treatment for epilepsy in pregnancy: neurodevelopmental outcomes in the child. Cochrane Database Syst Rev. 30 de octubre de 2014;(10):CD010236.

106. World Health Organization. mhGAP: scaling up care for mental, neurological and substance use disorders. Geneva: World Health Organization; 2011.

107. Edwards AGK, Naik G, Ahmed H, Elwyn GJ, Pickles T, Hood K, et al. Personalised risk communication for informed decision making about taking screening tests. Cochrane Database Syst Rev. 28 de febrero de 2013;(2):CD001865.

108. Royal College of Obstetrics and Gynaecology: Guidance for maternal medicine services in the coronavirus (COVID-19) pandemic Information for healthcare professionals. Version 2.5: Published Wednesday 9 December 2020.

109. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 18 de marzo de 2020;m441.

110. Varela L, López-Menéndez J, Redondo A, Fajardo ER, Miguelena J, Centella T, et al. Mortality risk prediction in infective endocarditis surgery: reliability analysis of specific scores. Eur J Cardio-Thorac Surg Off J Eur Assoc Cardio-Thorac Surg. 01 de 2018;53(5):1049-54.

111. Wang TKM, Wang MTM, Pemberton J. Risk scores and surgery for infective endocarditis: A meta-analysis. Int J Cardiol. 1 de noviembre de 2016;222:1001-2.

112. Pivatto Júnior F, Bellagamba CC de A, Pianca EG, Fernandes FS, Butzke M, Busato SB, et al. Análise de Escores de Risco para Predição de Mortalidade em Pacientes Submetidos à Cirurgia Cardíaca por Endocardite. Arq Bras Cardiol [Internet]. 2020 [citado 12 de mayo de 2021]; Disponible en: https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0066-782X2020000300518

113. Gatti G, Sponga S, Peghin M, Givone F, Ferrara V, Benussi B, et al. Risk scores and surgery for infective endocarditis: in search of a good predictive score. Scand Cardiovasc J. 4 de mayo de 2019;53(3):117-24.

114. Varela Barca L, Navas Elorza E, Fernández-Hidalgo N, Moya Mur JL, Muriel García A, Fernández-Felix BM, et al. Prognostic factors of mortality after surgery in infective endocarditis: systematic review and meta-analysis. Infection. diciembre de 2019;47(6):879-95.

115. Varela Barca L, Fernández-Felix BM, Navas Elorza E, Mestres CA, Muñoz P, Cuerpo-Caballero G, et al. Prognostic assessment of valvular surgery in active infective endocarditis: multicentric nationwide validation of a new score developed from a meta-analysis. Eur J Cardiothorac Surg. 1 de abril de 2020;57(4):724-31.

116. De Feo M, Cotrufo M, Carozza A, De Santo LS, Amendolara F, Giordano S, et al. The Need for a Specific Risk Prediction System in Native Valve Infective Endocarditis Surgery. Sci World J. 2012;2012:1-8.

117. Debray TPA, Koffijberg H, Nieboer D, Vergouwe Y, Steyerberg EW, Moons KGM. Meta-analysis and aggregation of multiple published prediction models: Meta-analysis and aggregation of multiple published prediction models. Stat Med. 30 de junio de 2014;33(14):2341-62.

118. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

119. Frank E Harrell Jr (2018). rms: Regression Modeling Strategies. R package version 5.1-2. https://CRAN.R-project.org/package=rms.

Apéndice

8. APÉNDICE

APÉNDICE 1: PROGNOSTIC MODELS FOR MORTALITY AFTER CARDIAC SURGERY IN PATIENTS WITH INFECTIVE ENDOCARDITIS: PROTOCOL.

APÉNDICE 2: PROGNOSTIC MODELS FOR MORTALITY AFTER CARDIAC SURGERY IN PATIENTS WITH INFECTIVE ENDOCARDITIS: SUPPLEMENTARY MATERIAL.

APÉNDICE 3: PROGNOSTIC FACTORS OF MORTALITY AFTER SURGERY IN INFECTIVE ENDOCARDITIS: SYSTEMATIC REVIEW AND META –ANALYSIS

APÉNDICE 4: PROGNOSTIC ASSESSMENT OF VALVULAR SURGERY IN ACTIVE INFECTIVE ENDOCARDITIS: MULTICENTRIC NATIONWIDE VALIDATION OF A NEW SCORE DEVELOPED FROM A META-ANALYSIS

APÉNDICE 5: PROTOCOL FOR DEVELOPMENT AND VALIDATION OF A CLINICAL PREDICTION MODEL FOR ADVERSE PREGNANCY OUTCOMES IN WOMEN WITH GESTATIONAL DIABETES

APÉNDICE 6: LECTURA CRÍTICA DE REVISIONES SISTEMÁTICAS DE ESTUDIOS DE PRONÓSTICO O RIESGO

Prognostic models for mortality after cardiac surgery in patients with infective endocarditis: a systematic review

Borja Manuel Fernandez-Felix, Laura Varela Barca, Esther García-Esquinas, Alfonso Muriel, Jesús López-Alcalde, Andrea Correa-Pérez, José Ignacio Pijoan, Aida Ribera, Josep Ramón Marsal, Marta Roqué, Nuria Fernández-Hidalgo, Javier Zamora

Citation

Borja Manuel Fernandez-Felix, Laura Varela Barca, Esther García-Esquinas, Alfonso Muriel, Jesús López-Alcalde, Andrea Correa-Pérez, José Ignacio Pijoan, Aida Ribera, Josep Ramón Marsal, Marta Roqué, Nuria Fernández-Hidalgo, Javier Zamora. Prognostic models for mortality after cardiac surgery in patients with infective endocarditis: a systematic review. PROSPERO 2020 CRD42020192602 Available from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020192602

Review question

We aim to systematically review all predicting models of post-operative mortality in patients undergoing cardiac surgery for active infective endocarditis models.

If sufficient homogenous and overlapping models are found, we plan to develop a meta-model to estimate mortality risk in these patients.

Searches

We will conduct a literature search to identify all potential studies for inclusion. We will apply no restriction considering language or publication dates nor status.

We will use the methodologic filter developed by Geersing et al. for prediction models research in MEDLINE and we will adapt it to use in EMBASE.

The sources used in the search will include MEDLINE (via Ovid) and EMBASE.

The exact search strategy is attached

Types of study to be included

According to the CHARMS checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies, the following types of studies will be included:

• Prognostic model development studies without external validation with independent data.

• Prognostic model development studies with external validation with independent data.

• Validation studies updating or extending previously published models.

Condition or domain being studied

Infective endocarditis treated with cardiac surgery.

Participants/population

We will include studies on adult patients (aged ? 18 years) undergoing cardiac surgery for infective endocarditis.

Intervention(s), exposure(s)

We will assess multivariable prognostic models developed, with or without external validation, and update or extend models in validation studies to predict the risk of post-operative mortality in patients undergoing cardiac surgery for active infective endocarditis.

Comparator(s)/control
Not applicable. Given the prognostic nature of this systematic review we do not have a comparator/control group.

Main outcome(s)
We will include models that predict early mortality after cardiac surgery (mortality 30 days after surgery or in-hospital mortality after surgery). We will admit mortality as defined by the authors.

* Measures of effect

Regression coefficient

Additional outcome(s)
No additional outcomes.

* Measures of effect

None

Data extraction (selection and coding)
We will identify and exclude duplicates. Two review authors will independently screen the results of the search strategies for eligibility by reading the titles and abstracts. In the case of disagreement we will ask a third review author to resolve disagreements. We will retrieve the full?text study reports of eligible studies, and two review authors will independently assess them for final inclusion in the review. We will resolve any disagreements through discussion and again, in the case of disagreement, we will ask a third review author to solve disagreements.

We will document the search and selection process in a flow chart as recommended in the Preferred Reporting Items for Systematic Reviews and Meta?Analyses (PRISMA) statement. Two review authors will independently extract data according to the CHARMS checklist. We will contact authors of individual studies for additional information, if required. We will use a standardised data extraction form containing the following sections:

• General information

• Source of data

• Participants (clinical and demographic data, eligibility and recruitment criteria)

• Outcome(s) to be predicted (type of outcome, definition, timing)

• Candidate predictors (number and type of predictors, definitions and handling)

• Sample size (number of participants, number of outcomes/events, number of events per variable)

• Missing data (number of participants with missing values and methods applied for handling missing data)

• Model development (modelling strategy, method used to select predictors, and shrinkage methods)

• Model performance statistics (calibration and discrimination)

• Model evaluation (methods used for internal and/or external validation)

• Results (final models presented, alternative presentation of the final model)

Data obtained from each eligible study will be summarized by providing descriptive tables reporting author name, publication year, modelling method, outcome, sample size, number of events, number of predictors included in the model, method for selection of predictors, type of model validation and measures of model performance.

Risk of bias (quality) assessment

We will use the 'PROBAST tool – A risk?of?bias tool for prediction?modelling studies'- to evaluate risk of bias. We will use the relevant items for development only, and for development and validation studies, to assess risk of bias on the following domains: Participants, predictors, outcome and analysis.

Items within each domain will be answered as yes (Y), probably yes (PY), no (N), probably no (PN), or no information (NI). A “yes” answer will indicate low risk of bias, while a “no” will indicate high risk of bias. The domains will be classified in:

• 'Low risk': the criterion is adequately fulfilled in the study.

• 'High risk': the criterion is not fulfilled in the study.

• 'Unclear': if the study report does not provide enough information to allow for a clear judgement, or if the risk of bias is unknown for one of the criteria listed above.

To note that a “no information” answer isn’t necessary indicative of bias.

On the basis of the risk of bias classification for each domain, authors will judge the overall risk of bias of the prediction model as low, moderate or high.

A summary of the risk of bias assessment will be reported by means of a table and/or graph.

Strategy for data synthesis

If we collect enough quality data on overlapping and homogenous sets of predictors, we will combine them into a so-called “meta-model”.

We will use an individual patient data set collected from the multicentre nationwide GAMES cohort (Grupo de Apoyo al Manejo de la Endocarditis infecciosa en España) to drive the weighting process of individual predictors to estimate the meta-model. GAMES cohort includes patients from 32 different hospitals across Spain. GAMES is active from January 1, 2008 and prospectively enrols patients in a nationwide registry. Regional and local ethics committees approved the study, and patients gave their informed consent.

We will extract the coefficients of the original models included in the review. We will assess individual model performance applying the original model to the patients included in the cohort described above.

Each model performance will be appraised in terms of calibration and discrimination and from this assessment a weight will be estimated for all coefficients in the model. The final step will be to aggregate individual predictors in the meta-model by a weighted average.

We will provide a summary table reporting the coefficients from original, updated and aggregated models along with their predictive performance.

Analysis of subgroups or subsets

If sufficient data is available, we will investigate and discuss clinical and methodological sources of heterogeneity. We will consider as sources of heterogeneity the following:

a) Methodological sources:

• Modelling methods.

• Time of event or event occurrence.

• Definitions and handling of continuous predictors.

b) Clinical sources:

• Only left or Left and right site IE

• Native or prosthetic valve IE

If the number of studies is appropriate, we will undertake the following sensitivity analyses:

• We will exclude studies with high or moderate risk of bias.

Contact details for further information

Borja Manuel Fernandez-Felix
borjamanuel86@gmail.com

Organisational affiliation of the review

Hospital Universitario Ramón y Cajal. IRYCIS. CIBER Epidemiology and Public Health.
https://www.ciberisciii.es/

Review team members and their organisational affiliations

Mr Borja Manuel Fernandez-Felix. Clinical Biostatistics Unit. Hospital Universitario Ramon y Cajal (IRYCIS), Madrid, Spain. CIBER Epidemiology and Public Health (CIBERESP), Madrid, Spain.
Miss Laura Varela Barca. Department of Cardiovascular Surgery. Hospital Universitario Son Espases, Palma de Mallorca, Spain.
Dr Esther García-Esquinas. CIBER Epidemiology and Public Health (CIBERESP), Madrid, Spain. Department of Preventive Medicine and Public Health, Universidad Autónoma de Madrid and Idipaz, Madrid, Spain.
Dr Alfonso Muriel. Clinical Biostatistics Unit. Hospital Universitario Ramon y Cajal (IRYCIS), Madrid, Spain. CIBER Epidemiology and Public Health (CIBERESP), Madrid, Spain.
Mr Jesús López-Alcalde. Clinical Biostatistics Unit. Hospital Universitario Ramon y Cajal (IRYCIS), Madrid, Spain. Faculty of Health Sciences, Universidad Francisco de Vitoria (UFV)-Madrid, Madrid, Spain. Cochrane Associate Centre of Madrid, Madrid, Spain.
Miss Andrea Correa-Pérez. Clinical Biostatistics Unit. Hospital Universitario Ramon y Cajal (IRYCIS), Madrid, Spain. Faculty of Health Sciences, Universidad Francisco de Vitoria (UFV)-Madrid, Madrid, Spain. Cochrane Associate Centre of Madrid, Madrid, Spain.
Dr José Ignacio Pijoan. CIBER Epidemiology and Public Health (CIBERESP), Madrid, Spain. Research and Methodological Support Unit, Hospital Universitario Cruces, Bilbao, Spain.
Dr Aida Ribera. Unitat d'Epidemiologia, Servei de Cardiologia, Hospital Universitari Vall d'Hebron, Barcelona, Spain. CIBER Epidemiology and Public Health (CIBERESP).
Josep Ramón Marsal. Unitat d'Epidemiologia, Servei de Cardiologia, Hospital Universitari Vall d'Hebron, Barcelona, Spain. CIBER Epidemiology and Public Health (CIBERESP).
Dr Marta Roqué. Iberoamerican Cochrane Centre, Biomedical Research Institute Sant Pau (IIB Sant Pau). CIBER Epidemiology and Public Health (CIBERESP). Centro Barcelona.
Dr Nuria Fernández-Hidalgo. Department of Infectious Diseases, Hospital Universitari Vall d'Hebron, Barcelona, Spain.
Dr Javier Zamora. Clinical Biostatistics Unit. Hospital Universitario Ramon y Cajal (IRYCIS), Madrid, Spain. CIBER Epidemiology and Public Health (CIBERESP), Madrid, Spain. Barts Research Centre for Women's Health, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom.

Type and method of review

Prognostic, Systematic review

Anticipated or actual start date

01 June 2020

Anticipated completion date

28 February 2021

Funding sources/sponsors

Centro de Investigación Biomédica en Red (CIBER)

Conflicts of interest
The lead reviewer (BFF) declares that he has no conflicts of interest. The review team has no conflicts of interest to disclose.
None known

Language
English

Country
Spain

Stage of review
Review Ongoing

Subject index terms status
Subject indexing assigned by CRD

Subject index terms
Cardiac Surgical Procedures; Endocarditis; Endocarditis, Bacterial; Humans; Prognosis

Date of registration in PROSPERO
12 August 2020

Date of first submission
16 June 2020

Stage of review at time of this submission

Stage	Started	Completed
Preliminary searches	Yes	No
Piloting of the study selection process	No	No
Formal screening of search results against eligibility criteria	No	No
Data extraction	No	No
Risk of bias (quality) assessment	No	No
Data analysis	No	No

The record owner confirms that the information they have supplied for this submission is accurate and complete and they understand that deliberate provision of inaccurate information or omission of data may be construed as scientific misconduct.

The record owner confirms that they will update the status of the review when it is completed and will add publication details in due course.

Versions
12 August 2020

Supplementary material

S1: Search strategies

The following exact search was used (search date 01/06/2020):

Ovid (Medline)

1. exp Endocarditis/
2. endocarditi*.tw.
3. 1 or 2
4. Cardiac Surgical Procedures/
5. (cardiac and (surger* or procedure*)).tw.
6. 4 or 5
7. 3 and 6
8. Validat$.af.
9. Predict$.ti.
10. Rule$.af.
11. 8 or 9 or 10
12. (Predict$ and (Outcome$ or Risk$ or Model$)).af.
13. ((History or Variable$ or Criteria or Scor$ or Characteristic$ or Finding$ or Factor$) and (Predict$ or Model$ or Decision$ or Identif$ or Prognos$)).af.
14. Decision$.af.
15. Logistic Models/
16. Model$.af.
17. Clinical$.af.
18. 15 or 16 or 17
19. 14 and 18
20. (Prognostic and (History or Variable$ or Criteria or Scor$ or Characteristic$ or Finding$ or Factor$ or Model$)).af.
21. 11 or 12 or 13 or 19 or 20
22. exp ROC Curve/
23. stratification.af.
24. discrimination.af.
25. discriminate.af.
26. c-statistic.af.
27. c statistic.af.
28. "Area under the curve".af.
29. AUC.af.
30. calibration.af.
31. indices.af.
32. algorithm.af.
33. multivariable.af.
34. 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33
35. 21 or 34
36. 7 and 35

Embase (Elsevier)

#1 'endocarditis'/exp

#2 endocardit*:ab,ti

#3 #1 OR #2

#4 'heart surgery'/exp

#5 cardiac:ab,ti AND (surger*:ab,ti OR procedure*:ab,ti)

#6 #4 OR #5

#7 #3 AND #6

#8 validat*:ab,ti

#9 predict*:ti

#10 rule*:ab,ti

#11 #8 OR #9 OR #10

#12 predict*:ab,ti AND (outcome*:ab,ti OR risk*:ab,ti OR model*:ab,ti)

#13 (history:ab,ti OR variable*:ab,ti OR criteria:ab,ti OR scor*:ab,ti OR characteristic*:ab,ti OR finding*:ab,ti OR factor*:ab,ti) AND (predict*:ab,ti OR model*:ab,ti OR decision*:ab,ti OR identif*:ab,ti OR prognos*:ab,ti)

#14 decision*:ab,ti

#15 'statistical model'/exp

#16 model*:ab,ti

#17 clinical*:ab,ti

#18 #15 OR#16 OR #17

#19 #14 AND #18

#20 prognostic:ab,ti AND (history:ab,ti OR variable*:ab,ti OR criteria:ab,ti OR scor*:ab,ti OR characteristic*:ab,ti OR finding*:ab,ti OR factor*:ab,ti OR model*:ab,ti)

#21 #11 OR #12 OR #13 OR #19 OR #20

#22 'receiver operating characteristic'/exp

#23 stratification:ab,ti

#24 discrimination:ab,ti

#25 discriminate:ab,ti

#26 'c-statistic':ab,ti

#27 'c statistic':ab,ti

#28 'area under the curve':ab,ti

#29 auc:ab,ti

#30 calibration:ab,ti

#31 indices:ab,ti

#32 algorithm:ab,ti

#33 multivariable:ab,ti

#34 #22 OR #23 OR #24 OR #25 OR #26 OR #27 OR #28 OR #29 OR #30 OR #31 OR #32 OR #33

#35 #21 OR #34

#36 #7 AND #35 #37 #7 AND #35 AND ([embase]/lim OR [pubmed-not-medline]/lim)

S2: Data extraction

Information on the following items was extracted using a standardized form based on CHARMS (CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies):

1. Study information: Author, year, journal and model´s name.
2. Source of data.
3. Participants: Recruitment method and dates; study setting; study regions and number of centers involved; inclusion and exclusion criteria; patient´s age (mean and standard deviation or median and interquartile range); number and percentage of native valve endocarditis; number, percentage and type (i.e. aortic, mitral, pulmonary or tricuspid) of valves affected.
4. Outcome: Definition and timing of occurrence.
5. Predictors: Number of candidate predictors; type of predictors; definition; and timing of measurement (preoperative or intraoperative)
6. Analysis:
    a. Sample size: Number of participants, events and events per predictor/parameter (EPP).
    b. Missing data: Number of participants with any missing value and methods used to handle missing data.
    c. Model development: Modelling method; method for selection of candidate predictors; method for selection of predictors during multivariable modelling
    d. Model performance: Discrimination and calibration measures.
    e. Model evaluation: Type of validation (apparent, internal or external) and optimism adjustment.
    f. Model results: Number of predictors included in the final model; presentation (e.g. coefficients and confidence interval); inclusion of model´s constant; alternative presentation of the final model.

S3: Critical appraisal and applicability

Model were assessed to risk of bias using a standardized form based on the PROBAST on the following domains: Participants; Predictors; Outcome; Analysis.

The signalling questions were answered for each domain with one out of these options ('yes', 'probably yes', 'probably no', 'no', 'no information'); where 'yes' means the absence of a potential bias. We rated domain-level 'Risk of bias' assessments as:

- Low risk of bias: if the criterion is adequately fulfilled in the study, i.e. the study is at a low risk of bias for the given domain.
- High risk of bias: if the criterion is not fulfilled in the study, i.e. the study is at high risk of bias for the given domain.
- Unclear risk of bias: if the study report does not provide enough information to allow for a clear judgement or if the risk of bias is unknown for one of the domains listed above.

The applicability judgement of the model to the research question occurs per following domains: Participants, Predictors and Outcome. The possible responses were: 'low concern regarding applicability', 'high concern regarding applicability' and 'unclear concern regarding applicability' (equivalent to the categories for risk of bias).

If risk of bias or applicability were high in at least one of the domains, overall risk of bias or applicability was judged high. If at least one of the answers was “No” or “Probably no,” the judgment could still be low risk of bias, in this case specific reasons were provided. The complete information about of the 'Risk of bias' and 'Applicability' assessment of the authors is shown in **Supplementary Table S8 and Figure S1**.

S4: Data imputation

We used linear regression imputation for continuous variables, truncated regression imputation for continuous variable with a restricted range, logistic regression imputation for binary data, multinomial logistic regression imputation for unordered categorical data and ordered logistic regression imputation for ordered categorical data.

S5: Statistical software

The analyses were conducted in Stata version 16 using mi command for multiple imputation, `mfpmi` command for estimation meta-model coefficients using logistic regression modelling in presence of multiple imputation datasets, `roctab` and `logistic` command for C-statistics, slope calibration and calibration-in-the-large calculations. These commands were combined in a syntax (available from the corresponding author upon reasonable request) to obtain bootstrap confidence intervals and performance measures adjusted for optimism. `Forestplot` and `pmcalplot` commands were used for figures.

Table S1: Characteristics of patients included in the validation dataset (GAMES registry)

	Mortality		Missing data
	No (n=1,099)	Yes (n=354)	
	n (%)	n (%)	n
Patient related-factors			
Age (years), mean(sd)	62.0 (13.4)	68.9 (10.0)	-
Female	275 (25.1%)	112 (31.8%)	6
Chronic pulmonary disease	179 (18.3%)	83 (26.9%)	165
Diabetes	248 (22.6%)	131 (37.0%)	2
Hypertension	546 (49.8%)	238 (67.4%)	4
Pulmonary hypertension	58 (5.3%)	27 (7.6%)	-
Creatinine (mg/dl.), mean(sd)	1.1 (0.9)	1.4 (1.1)	56
Prior CABG	56 (5.1%)	31 (8.8%)	4
Prior valvular surgery	356 (32.5%)	168 (47.6%)	6
LVEF (%), mean(sd)	59.8 (11.0)	58.0 (12.0)	366
Clinical presentation related-factors			
Septic shock	85 (7.7%)	102 (28.8%)	
NYHA			25
• I	883 (81.5%)	241 (69.9%)	
• II	158 (14.6%)	68 (19.7%)	
• III	30 (2.8%)	27 (7.8%)	
• IV	12 (1.1%)	9 (2.6%)	
Preoperative status			23
• Elective	746 (69.1%)	180 (51.3%)	
• Urgent	265 (24.6%)	115 (32.8%)	
• Emergent	68 (6.3%)	56 (16.0%)	
Valves affected			-
• 0	13 (1.2%)	3 (0.8%)	
• 1	913 (83.1%)	288 (81.4%)	
• 2	169 (15.4%)	60 (16.9%)	
• 3	4 (0.4%)	3 (0.8%)	
Surgery-related factors			

Abscess	284 (26.0%)	124 (35.3%)	8
Fistula	34 (3.1%)	23 (6.5%)	-
Dehiscence	117 (10.7%)	63 (17.8%)	2
Weight of intervention			-
• Single non-CABG	867 (78.9%)	273 (77.1%)	
• 2 procedures	225 (20.5%)	74 (20.9%)	
• 3 procedures	7 (0.6%)	7 (2.0%)	
Surgery in aorta	24 (2.2%)	11 (3.1%)	-
IE-related factors			
Type of valve			15
• Natural	754 (69.4%)	186 (52.8%)	
• Prosthetic	332 (30.6%)	166 (47.2%)	
Valve location			
• No valve treated	13 (1.2%)	3 (0.8%)	
• Aortic	547 (49.8%)	164 (46.3%)	
• Mitral	350 (31.8%)	121 (34.2%)	
• Pulmonary	2 (0.2%)	0 (0.0%)	
• Tricuspid	14 (1.3%)	3 (0.8%)	
• Multiple	173 (15.7%)	63 (17.8%)	
Infection etiology			52
• *Staphylococcus* spp.	367 (34.7%)	190 (55.2%)	
- coagulase-negative staphylococci	208 (57%)	92 (48%)	
- S. *aureus*	159 (43%)	98 (52%)	
MSSA	115	75	
MIRSA	0	2	
MRSA	23	12	
Unknown	21	9	
• *Pseudomonas* spp.	3 (0.3%)	4 (1.2%)	
• Fungal disease	20 (1.9%)	10 (2.9%)	
• *Streptococcus* spp.	363 (34.3%)	70 (20.3%)	
• Other microorganisms	304 (28.8%)	70 (20.3%)	

n: number of patients; sd: standard deviation; CABG: coronary artery bypass grafting; LVEF: left ventricular ejection fraction; NYHA: New York Heart Association; MSSA: methicillin sensitivity S. *aureus*; MIRSA: methicillin intermediate resistant S. *aureus*; MRSA: methicillin resistant S. *aureus*

Table S2: Studies excluded and motive of exclusion

DOI / PMID	Reference
Medically treated patients	
PMID: 3893114	Alsip S G, Blackstone E H, Kirklin J W, Cobbs C G. 1985. "Indications for cardiac surgery in patients with active infective endocarditis". *The American journal of medicine* 78(6B):138-48.
PMID: 2759756	Woo K S, Lam Y M, Kwok H T, Tse L K. K, Vallance-Owen J. 1989. "Prognostic index in prediction of mortality from infective endocarditis". *International Journal of Cardiology* 24(1):47-54.
	Kjaergaard J, Rasmussen R, Bruun N, Hassager C. 2009. "Vegetation length or area: Which is the better predictor of outcome in infective endocarditis?". *International Journal of Antimicrobial Agents* 33:S27-S28.
10.1177/2048872615574706	Guimaraes ” Baseline predictors of in-hospital mortality in patients with infective endocarditis”. Abstracts for the Cardiac Society of Australia and New Zealand Annual Scientific Meeting and the International Society for Heart Research Australasian Section Annual Scientific Meeting. 2016. *Heart Lung and Circulation* 25:.
10.1016/j.recesp.2020.04.010	García-Granja P E, López J, Vilacosta I, Sarriá C, Domínguez F, Ladrón R, et al.. 2020. "Predictive model of in-hospital mortality in left-sided infective endocarditis". *Revista Espanola de Cardiologia* :.
10.1590/s0102-76382007000200007	Costa MA, Wollmann DR Jr, Campos AC, Cunha CL, Carvalho RG, Andrade DF, et al.. 2007. "Risk index for death by infective endocarditis: a multivariate logistic model.". *Revista brasileira de cirurgia cardiovascular : orgao oficial da Sociedade Brasileira de Cirurgia Cardiovascular* 22(2):192-200.
Did not provide a prognostic model	
10.1177/2048872616663431	Garcia Granja, P E, Lopez J, Ladron R, Vilacosta I, Olmos C, Ortiz Bautista, et al.. 2016. "Influence of valve culture in prognosis of leftsided infective endocarditis". *European Heart Journal: Acute Cardiovascular Care* 5:384-385.
10.1093/ejcts/ezv223	Patrat-Delon Solene, Rouxel Adrien, Gacouin Arnaud, Revest Matthieu, Flecher Erwan, Fouquet Olivier, et al.. 2016. "EuroSCORE II underestimates mortality after cardiac surgery for infective endocarditis". *European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery* 49(3):944-51.
10.1016/j.jescts.2017.02.004	Elmasry A, Omran A M, Elprince A, Elameen S, Mansy M M, Mahlab A S. 2017. "Predictors of in-hospital mortality in surgically treated valvular infective endocarditis cases at National Heart institute, Egypt". *Journal of the Egyptian Society of Cardio-Thoracic Surgery* 25(1):35-44.
10.1177/0218492318798258	Nagy Mohamad, Alkady Hesham, Abo Senna, Waleed, Abdelhay Soliman. 2018. "Predictors of surgical outcome in isolated prosthetic mitral valve endocarditis". *Asian cardiovascular & thoracic annals* 26(7):517-523.
10.1016/j.repc.2019.08.009	Guiomar N, Vaz-da-Silva M, Mbala D, Sousa-Pinto B, Monteiro J P, Ponce P,et al.. 2020. "Cardiac surgery in infective endocarditis and predictors of in-hospital mortality". *Revista Portuguesa de Cardiologia* :.
Congress abstract of a study included in full text review	
Original study ref: 10.1016/j.ijcard.2014.04.266	Martinez-Selles M, Munoz P, Arnaiz A, Moreno M, Galvez J, Rodriguez-Roda J, et al.. 2014. "Valve surgery in active infective endocarditis: A simple score to predict in-hospital prognosis". *European Heart Journal* 35:756.
Original study ref: 10.1093/icvts/ivv304	Madeira S, Santos M, Rodrigues R, Tralhao A, Mesquita J, Carmo J, et al.. 2015. "Assessment of operative mortality risk in patients with active infective endocarditis undergoing cardiac surgery: Performance of the EuroScore I and II logistic models". *European Heart Journal* 36:268.
Original study ref: 10.1136/heartjnl-2016-311093	Olmos C, Vilacosta I, Fernandez C, Tirado G, Freitas-Ferraz A, Lopez J, et al.. 2015. "Development and validation of a risk score for cardiac surgery in infective endocarditis". *European Heart Journal* 36:374.
Original study ref: 10.1007/s00380-014-0472-0	Wang T K. M, Oh T, Voss J, Kang N, Pemberton J. 2013. "Comparison and implications of contemporary risk scores for predicting mortality and morbidity after surgery for active infective endocarditis". European Heart Journal 34:502.
Validation studies without updating	

10.1007/s00380-014-0472-0	Wang Tom Kai Ming, Oh Timothy, Voss Jamie, Gamble Greg, Kang Nicholas, Pemberton James. 2015. "Comparison of contemporary risk scores for predicting outcomes after surgery for active infective endocarditis". Heart and vessels 30(2):227-34.
10.1080/14017431.2019.1610188	Gatti Giuseppe, Sponga Sandro, Peghin Maddalena, Givone Filippo, Ferrara Veronica, Benussi Bernardo, et al.. 2019. "Risk scores and surgery for infective endocarditis: in search of a good predictive score". Scandinavian cardiovascular journal : SCJ 53(3):117-124
Provided a composite outcome	
10.1001/jama.289.15.1933	Hasbun R, Vikram H R, Barakat L A, Buenconsejo J, Quagliarello V J. 2003. "Complicated Left-Sided Native Valve Endocarditis in Adults: Risk Classification for Mortality". *Journal of the American Medical Association* 289(15):1933-1940.
10.1136/hrt.2010.200295	Lopez Javier, Fernandez-Hidalgo Nuria, Revilla Ana, Vilacosta Isidre, Tornos Pilar, Almirante Benito, et al.. 2011. "Internal and external validation of a model to predict adverse outcomes in patients with left-sided infective endocarditis". *Heart (British Cardiac Society)* 97(14):1138-42.
Patients not diagnosed with infective endocarditis	
10.1016/j.ejcts.2011.01.002	Akar Ahmet Ruchan, Kurtcephe Murat. Sener Erol, Alhan Cem, Durdu Serkan, Kunt Ayse Gul, et al.. 2011. "Validation of the EuroSCORE risk models in Turkish adult cardiac surgical population". *European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery* 40(3):730-5.
Editorials and Coments	
Editorial: 10.1001/jama.289.15.1991	Granowitz E V, Longworth D L. 2003. "Risk Stratification and Bedside Prognostication in Infective Endocarditis". Journal of the American Medical Association 289(15):1991-1993.
Editorial: 10.36660/abc.20200070	Martins A B. B, Lamas C D. C. 2020. "Prognostic scores for mortality in cardiac surgery for infective endocarditis". Arquivos Brasileiros de Cardiologia 114(3):525-529.
Editorial: 10.1053/j.jvca.2018.02.005	Stein Erica, Andritsos Michael. 2018. "Risk Stratification and Optimization of Cardiac Surgical Patients With Infective Endocarditis: Does It Matter?". Journal of cardiothoracic and vascular anesthesia 32(6):2537-2539.
Editorial: 10.21037/jtd.2019.09.69	Tattevin Pierre, Fillatre Pierre, Tchamgoue Serge, Lesouhaitier Mathieu, Nesseler Nicolas, Tadie Jean-Marc. 2019. "Should we include microorganisms in scores to predict outcome in candidates for cardiac surgery during the acute phase of endocarditis?". Journal of thoracic disease 11(10):E158-E162.
Comment: 10.2169/internalmedicine.3579-19	Toyoda S, Saito F, Inoue T. 2020. "Authors' reply: How to construct novel criteria for predicting complication with infectious endocarditis". Internal Medicine 59(1):147-148.
Comment: 10.1016/j.ijcard.2015.08.167	Wang T K. M. 2016. "Risk scores for endocarditis surgery: Callout for reporting logistic models". International Journal of Cardiology 202:960.
Model built using a systematic review	
10.1093/ejcts/ezz328	Varela Barca, L, Fernández-Felix B M, Navas Elorza E, Mestres C A, Muñoz P, Cuerpo-Caballero G, et al.. 2020. "Prognostic assessment of valvular surgery in active infective endocarditis: Multicentric nationwide validation of a new score developed from a meta-analysis". *European Journal of Cardio-thoracic Surgery* 57(4):724-731.

92

Table S3: Characteristics of the primary studies.

Author, Year	Enrolment period	Study setting	Study design	Study region (Centers)	Age Mean (sd) or median (Q_1;Q_3)	Native valve (%)	Valves affected
In-hospital or 30 days mortality							
De Feo, 2012	1980 - 2009	Cardiac surgery centers	Retrospective cohort	Italy (1)	49 (16)	100	All
Gaca, 2011	2002 - 2008	Cardiac surgery centers	Registry (STS ACSD)	North America (Unclear)	55 (45;66)	NI	All
Madeira, 2016	2007 - 2014	Cardiac surgery centers	Retrospective cohort	Portugal (1)	60 (47;70)	73.4	All
In-hospital mortality							
Gatti, 2017a	2000-2015 (Italy) 2008 (France)	Cardiac surgery centers	Retrospective cohort and registry (AEPEI)	Italy (1) France (7)	59.1 (15.4)	78.9	All
Gatti, 2017b	1999 - 2015	Cardiac surgery centers	Retrospective cohort	Italy (1)	60.6 (8.5)	74.6	All
Martínez-Sellés, 2014	2008 - 2010	Cardiac surgery centers	Registry (GAMES)	Spain (26)	61.4 (15.5)	61.1	All
Olmos, 2017	1996 - 2014	Cardiac surgery centers	Retrospective cohort	Spain (3)	62 (14)	61.1	A/M
30 days mortality							
Di Mauro, 2017	2000 - 2015	Cardiac surgery centers	Retrospective cohort	Italy (26)	59.6 (15.1)	81.8	All
Fernández-Hidalgo, 2018	2000 - 2011	Cardiac surgery centers	Retrospective cohort	Spain (9)	58 (15.1)	NI	All

Sd: Standard deviation; Q_1: First quartil; Q_3: Thrird quartil: STS ACSD: The Society of Thoracic Surgeons Adult Cardiac Surgery Database; AEPEI: Association pour l'Etude et la Prevention de l'Endocadite Infectieuse; GAMES: Grupo de Apoyo al Manejo de la Endocarditis infecciosa en España; A: Aortic valve; M: Mitral valve; NI: No information.

Table S4: Definition of the predictors

Preoperative patient-related factors	
Age	**Di Mauro 2017; De Feo 2012; Fernández-Hidalgo 2018; Martínez-Sellés 2014; Olmos 2017; GAMES registry.**
Gender	**Di Mauro 2017; Martínez-Sellés 2014; GAMES registry.**
Renal failure	**Di Mauro 2017.** Creatinine ≥ 2 mg/dl. **Gaca 2011.** Documented history of renal failure and/or history of creatinine > 2 mg/dl. Prior renal transplant patients not included as pre-op renal failure unless since transplantation creatinine creatine values had been > 2.0 mg/dl. **De Feo 2012; GAMES registry.** Creatinine > 2 mg/dl. **Gatti 2017a.** eGFR <50 mL/min/1.73 m2. The creatinine clearance rate calculated according to the Cockcroft–Gault formula was used to estimate GFR. **Fernández-Hidalgo 2018.** Serum creatinine >200 mmol/l preoperatively. **Olmos 2017.** Renal failure was defined as GFR <60 mL/min/1.73 m2.
Body max index	**Gatti 2017a.**
Chronic pulmonary disease	**Di Mauro 2017.** Long term use of bronchodilators or steroids for lung disease. **Gaca 2011; GAMES registry.** Chronic lung disease.
Diabetes Mellitus	**Gaca 2011.** History of IDDM or NIDDM diabetes mellitus. Patients placed on a pre-operative diabetic pathway of Insulin drip but at admission were controlled with none, diet or oral method are not coded as insulin dependent.
Hypertension	**Gaca 2011.** Diagnosis of hypertension, documented by one of the following: a. Documented history of hypertension diagnosed and treated with medication, diet and/or exercise. b. Prior documentation of systolic blood pressure >140 mmHg or diastolic blood pressure > 90 mmHg for patients without diabetes or chronic kidney disease, or prior documentation of systolic blood pressure >130 mmHg or diastolic blood pressure > 80 mmHg on at least 2 occasions for patients with diabetes or chronic kidney disease. c. Currently on pharmacologic therapy to control hypertension.
Pulmonary hypertension	**Gatti 2017a.** Systolic pulmonary artery pressure > 55mmHg. **Fernández-Hidalgo 2018; GAMES registry.** Systolic pulmonary artery pressure > 60 mmHg.
Anemia	**Gatti 2017b.** Haemoglobin <12 g/dl for women and <13 g/dl for men.
Thrombocytopaenia	**Olmos 2017.** Platelet count <150.000/mL.
Left ventricular ejection fraction	**Di Mauro 2017; GAMES registry.** Percentage of left ventricular ejection fraction.
Arrhythmia	**Gaca 2011.** History of preoperative arrhythmia (sustained ventricular tachycardia, ventricular fibrillation, atrial fibrillation, atrial flutter, third degree heart block) treated with any of the following modalities: ablation therapy, AICD, pacemaker, pharmacological treatment or electrocardioversion.
Prior cardiac surgery	**Gaca 2011.** Prior CABG or prior valve surgery (i.e. previous surgical replacement and/or surgical repair of a cardiac valve, including percutaneous valve procedures). **Fernández-Hidalgo 2018; GAMES registry.** One or more previous major cardiac operations involving opening the pericardium.
Clinical presentation-related factors	
Critical preoperative state	**Di Mauro 2017; Gatti 2017a; Gatti 2017b; Fernández-Hidalgo 2018; GAMES registry.** Any one or more of the following: ventricular tachycardia or fibrillation or aborted sudden death, preoperative cardiac massage, preoperative ventilation before arrival in the anesthetic room, preoperative inotropic support, intra-aortic balloon counter pulsation or preoperative acute renal failure (anuria or oliguria,10 ml/h). **Gaca 2011.** Patient placed on IABP or received IV inotropic agents within 48 hours preceding surgery.

	De Feo 2012. (Ventilatory support in original paper) Patients admitted to the Cardiac Surgery Department on mechanical ventilation (intubated) or requiring ventilatory support by noninvasive ventilation during preoperative stay (generally for poor hemodynamic conditions and/or pulmonary edema). **Olmos 2017. (**Cardiogenic shock in original paper) Systolic pressure <90 mmHg and tissue hypoperfusion due to myocardial dysfunction, despite adequate preload, and accompanied by low cardiac index and high pulmonary wedge pressure.
NYHA functional class	**De Feo 2012; Gatti 2017a; Gatti 2017b; Fernández-Hidalgo 2018; GAMES registry.** NYHA classification for dyspnea: I: no symptoms on moderate exertion; II: symptoms on moderate exertion; III: symptoms on light exertion; IV: symptoms at rest.
Septic shock	**Olmos 2017.** Acute circulatory failure in sepsis, with persistent systolic pressure <90 mmHg despite adequate volume resuscitation.
EuroSCORE I	**Martínez-Sellés 2014.** European system for cardiac operative risk evaluation I. Nashef 1999.
EuroSCORE II	**Madeira 2016.** European system for cardiac operative risk evaluation II. Nashef 2011.
Surgery-related factors	
Paravalvular complications	**De Feo 2012.** Presence of either an annular abscess or aortocavitary fistula. **Di Mauro 2017.** Presence of an abscess. **Fernández-Hidalgo 2018.** Presence of a fistula. **Martínez-Sellés 2014.** (Substantial intracardiac destruction in original paper) Abscesses present or echocardiography findings suggestive of invasive infection (communication between chambers, wall dissection or large valvular dehiscence). **Olmos 2017.** Presence of abscess, pseudoaneurysm, fistula or prosthetic dehiscence. **GAMES registry.** purulent cavity with necrosis and capacity to invade adjacent structures.
Urgency of procedure	**Gaca 2011.** Urgent status: procedure required during the same hospitalization to minimize chance of further clinical deterioration; Emergency status: patient requiring emergency operations will have ongoing, refractory (difficult, complicated, and/or unmanageable) unrelenting cardiac compromise, with or without hemodynamic instability, and not responsive to any form of therapy except cardiac surgery. An emergency operation is one in which there should be no delay in providing operative intervention. **Fernández-Hidalgo 2018.** Urgent status: patients not electively admitted for operation but who require surgery on the current admission for medical reasons and cannot be discharged without a definitive procedure; Emergency status: operation before the beginning of the next working day after decision to operate. **Martínez-Sellés 2014.** Definition not available. **GAMES registry.** Urgent surgery: surgery required within 24 h of its indication; Emergency surgery: surgery required on the day of admission.
Number of treated valves/prostheses	**Di Mauro 2017; Gaca 2011; GAMES registry.** Number of treated valves/prostheses.
Weight of intervention	**Gatti 2017b.** Surgery on thoracic aorta.
IE-related factors	
Infection etiology	Pathogen isolated on blood or specimen culture. **Di Mauro 2017.** *Pseudomonas aeruginosa*; *Staphylococcus aureus*; Fungi; Other microorganisms. **Fernández-Hidalgo 2018; Martínez-Sellés 2014.** *Staphylococcus* spp. **Olmos 2017**. *Staphylococcus aureus* or fungi.

	GAMES registry. *Staphylococcus* spp. (coagulase-negative staphylococci or S. *aureus*); *Pseudomonas* spp.; Fungal disease; *Streptococcus* spp.; Other microorganisms.
Type of valve	**Madeira 2016; Olmos 2017.** Not available. **Martínez-Sellés 2014.** Prosthetic valve IE was defined as infection occurring on any type of non-native tissue or mechanical device.
Active endocarditis	**Gaca 2011** Type of endocarditis the patient has. If the patient is currently being treated for endocarditis, the disease is considered active. If no antibiotic medication (other than prophylactic medication) is being given at the time of surgery, then the infection is considered treated.
Valve location	**Fernández-Hidalgo 2018.** Infection location (aortic, mitral, other). **Games registry.** Infection location (aortic, mitral, pulmonary, tricuspid).
Positivity of latest pre-op. blood culture	**De Feo 2012.** Operation without possibility of previous attainment of negative cultures by antibiotic therapy (latest culture had always been performed within 5 to 7 days preoperatively).

eGFR: estimated glomerular filtration rate; **GFR**: glomerular filtration rate; **IDDM**: insulin-dependent diabetes mellitus; **NIDDM**: non-insulin-dependent; AICD: CABG: coronary artery bypass grafting; IABP: Intra-Aortic Balloon Pump; NYHA: New York Heart Association; **GAMES**: Grupo de Apoyo al Manejo de la Endocarditis infecciosa en España.

96

Table S5. Model compositions and percentage of missing data in GAMES registry.

	De Feo, 2012	Gaca, 2011	Madeira, 2016	Gatti, 2017a (Original)	Gatti, 2017a (Alternate)	Gatti, 2017b	Martínez-Sellés, 2014	Olmos, 2017	Di Mauro, 2017 (EndoSCORE)	Fernández-Hidalgo, 2018 (sp. ES-I)	Fernández-Hidalgo, 2018 (sp. ES-II)	Meta-model	Percentage of missing data in GAMES
Patient-related factor													
Renal failure	■	■		■	■			■	■	■		■	4%
Age (years)	■						■	■	■	■	■	■	0%
Prior cardiac surgery		■								■	■	■	1%
Gender							■		■			■	<1%
Chronic pulmonary disease		■							■			■	11%
Pulmonary hypertension				■						■		■	0%
Anemia						■							100%
BMI (kg/m)				■									29%
Diabetes Mellitus		■											<1%
Hypertension		■											<1%
Arrhythmia		■											<1%
Left ventricular ejection fraction (%)									■			■	7%
Thrombocytopaenia								■					100%
Clinical presentation-related factors													
Critical preoperative state	■	■		■	■	■		■	■	■	■	■	2%
NYHA functional class	■			■	■	■				■	■	■	2%
Septic shock								■					0%
EuroSCORE I							■						19%
EuroSCORE II			■										37%
Surgery-related factors													
Paravalvular complications	■						■	■	■	■	■	■	4%
Urgency of procedure		■					■			■	■	■	2%
Number of treated valves/ prostheses		■							■			■	0%
Weight of intervention						■							0%
IE-related factors													
Infection etiology							■	■	■	■	■	■	4%
Type of valve			■				■	■					1%
Valve location										■	■	■	0%
Active endocarditis		■											0%
Positivity of latest pre-op. blood culture	■												0%

RoB: Risk of Bias; GAMES: Grupo de Apoyo al Manejo de la Endocarditis infecciosa en España; BMI: body mass index; NYHA: New York Heart Association; IE: infective endocarditis; pre-op: pre-operative.
Dark cells indicate that the predictor was included in the model.

Table S6: Minimum sample size for development of a new multivariable prediction model.

Author, Year	Available data			Minimum Sample Size[a]/EPP required for development of a new multivariable prediction model		
	Events n (%)	Candidate predictors	Sample size/EPP	Explained variability scenarios		
				10%	20%	30%
De Feo, 2012	40 (9.1)	19	440 / 2.1	3,651 / 17.5	1,777 / 8.5	1,152 / 5.5
Gaca, 2011	1,117 (8.2)	38	13,617 / 29.4	7,709 / 16.6	3,757 / 8.1	2,439 / 5.3
Madeira, 2016	21 (16.4)	15	128 / 1.4	2,211 / 24.2	1,067 / 11.7	685 / 7.5
Gatti, 2017a (Original & Alternate)	56 (15.5)	57	361 / 1.0	8,589 / 23.4	4,147 / 11.3	2,664 / 7.2
Gatti, 2017b	28 (20.3)	56	138 / 0.5	7,649 / 27.7	3,679 / 13.3	2,353 / 8.5
Martínez-Sellés, 2014	106 (24.3)	NI	437 / NI	n.a.	n.a.	n.a.
Olmos, 2017	124 (29.2)	37	424 / 3.4	4,562 / 36.0	2,185 / 17.2	1,390 / 11.0
Di Mauro, 2017	298 (11.0)	32	2,715 / 9.3	5,600 /19.2	2,718 / 9.3	1,756 / 6.0
Fernández-Hidalgo, 2018 (Sp. ES-I)	208 (26.7)	26	779 / 8.0	3,277 / 33.6	1,571 / 16.1	1,001 / 10.3
Fernández-Hidalgo 2018 (Sp. ES-II)	208 (26.7)	27	779 / 7.7	3,403 / 33.6	1,631 / 16.3	1,039 / 10.3

Sp. ES-I: specific EuroSCORE I; Sp. ES-II: specific EuroSCORE II; n: number of events; EPP: events per parameter; NI: not informed; n.a.: not applicable.

[a] We calculated the minimum sample size required for the development of a new multivariable prediction model using the criteria proposed by Riley et al. (1). We used the number of candidate predictors and mortality rates from the original paper, and we considered three different scenarios for the variability explained by the model (10%, 20% or 30%). Prediction models with C-statistics between 0.7 and 0.8 typically have R-squared values between 10 and 20% (2) and were models which reported C-statistic close to 0.9. For a mortality proportion of 0.2. the max(R^2_{CS}) is 0.63 (1), therefore for the 10% explained variability scenario R^2_{CS} = 0.63*0.10 = 0.063.
R^2_{CS}: Cox-Snell R-squared

We used pmsampsize stata command developed by Riley R. and Ensor J.

1. Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Statistics in Medicine. 2019 Mar 30;38(7):1276–96.

2. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating [Internet]. 2019 [cited 2020 Apr 28]. Available from: https://doi.org/10.1007/978-3-030-16399-0

Table S7: Prognostic models equation

De Feo 2012	(No constant) 0.041xage + 1.076 (if renal failure) + 1.777 (if NYHA class IV) + 2.281 (if critical preoperative state) + 1.093 (if positivity of latest pre-op. blood culture) + 1.110 (if paravalvular complications)
Di Mauro 2017	-2.60 + 0.46 (if age 60-70y) + 0.88 (if age 70-80y) + 1.53 (if age>80y) + 0.51 (if female) – 0.03xLVEF + 0.50 (if renal failure) + 0.68 (if chronic pulmonary disease) + 1.46 (if critical preoperative state) + 0.50 (if two valves/prostheses treated) + 1.50 (if three valves/prostheses treated) + 1.09 (if paravalvular complications) + 1.46 (if *Pseudomonas aeruginosa*) + 1.24 (if *Staphylococcus aureus*) + 1.66 (if fungi) + 0.60 (if other microorganisms)
Fernández-Hidalgo 2018	**Specific ES-I:** -3.132 + 1.101 (if prior cardiac surgery) + 1.121 (if critical preoperative state) + 0.464 (if renal failure) + 0.702 (if NYHA class > 1) + 0.059x(age-60) (if age > 60y) + 0.806 (if emergency status) + 1.220 (if paravalvular complications) + 0.528 (if *Staphylococcus* spp.) – 1.268 (if pulmonary hypertension) + 0.374 (if mitral location) **Specific ES-II:** -4.210 + 0.964 (if prior cardiac surgery) + 1.024 (if critical preoperative state) + 0.617 (if NYHA class > 1) + 0.062x(age-60) (if age > 60y) + 1.950 (if emergency status) + 1.157 (if urgent status) + 1.141 (if paravalvular complications) + 0.531 (if *Staphylococcus* spp.) + 0.383 (if mitral location)
Gaca 2011	(No constant) 0.490 (if Prior CABG) + 0.422 (if urgent status) + 1.153 (if cardiogenic shock) + 0.672 (if critical preoperative state) + 0.602 (if multiple valve procedure) + 0.471 (if prior valve surgery) + 0.547 (if IDDM) + 0.431 (if NIDDM) + 0.342 (if hypertension) + 0.344 (if chronic pulmonary disease) + 0.695 (if active endocarditis) + 0.827 (if renal failure) + 0.504 (if arrhythmia)
Gatti 2017a	**Original:** -3.065 + 0.58 (if BMI > 27kg/m^2) + 1.26 (if renal failure) + 0.75 (if NYHA class IV) + 0.58 (if pulmonary hypertension) + 0.86 (if critical preoperative state) **Alternate:** -1.411 + 1.32 (if renal failure) + 0.75 (if NYHA class IV) + 0.85 (if critical preoperative state)
Gatti 2017b	**Preoperative:** (No constant) 2.40 (if anemia) + 0.96 (if NYHA class IV) + 1.60 (if critical preoperative state) + 1.86 (if paravalvular complications) + 2.02 (if surgery on thoracic aorta)
Madeira 2016	(No constant) 1.932 (if prosthetic valve IE) + 0.081xEuroSCORE-II
Martínez-Sellés 2014	(No constant) 0.030xage + 0.790 (if prosthetic valve IE) + 0.640 (if paravalvular complications) + 0.740 (if female) + 0.690 (if urgent status) + 0.830 (if *Staphylococcus* spp.) + 0.02xEuroSCORE-I
Olmos 2017	-3.358 + 0.916 (if age 52-63y) + 1.336 (if age 64-72y) + 1.362 (if age≥73y) + 0.645 (if prosthetic endocarditis) + 0.903 (if *Staphylococcus aureus* or fungi) + 0.702 (if septic shock) + 0.655 (if thrombocytopenia) + 0.542 (if renal failure) + 1.486 (if cardiogenic shock) + 0.541 (if paravalvular complications)

NYHA: New York Heart Association; LVEF: Left ventricular ejection fraction; CABG: Coronary artery bypass graft; IDDM: insulin-dependent diabetes mellitus; NIDDM: non-insulin-dependent diabetes mellitus; BMI: Body mass index: IE: Infective endocarditis.

Box S1: meta-model equation and example of use

The equation of the meta-model to estimate probability of mortality in patient with infective endocarditis is as follows:

$$P(mortality) = \frac{\exp(Y)}{1 + \exp(Y)}$$

where Y = −5.00 + 0.22 [if female] + 0.045 * age + 0.28 [if renal failure] + 0.51 [if prior cardiac surgery] + 0.29 [if chronic pulmonary disease] + 0.17 [if pulmonary hypertension] - (0.013 * LVEF) + 1.17 [if critical preoperative state] + 0.33 [if NYHA>I] + 0.43 [if abscess] + 0.59 [if fistulae] + 0.44 [if urgent status] + 0.85 [if emergency status] + 0.22 [if two valves treated] + 0.65 [if three valves treated] + 0.19 [if mitral location] + 0.64 [if *Staphylococcus spp.*] + 0.61 [if Fungi]

Example:

A 60-year-old woman with renal failure and pulmonary hypertension, with a left ventricular ejection fraction of 60%, NYHA-II, with paravalvular abscess. The preoperative condition is not critical, but the patient must undergo urgent surgery. Infective endocarditis is located in the aortic valve and was caused by *Staphylococcus spp*.

Y = −5.00 + 0.22 [female] + 0.045*60 + 0.28 [renal failure] + 0.17 [pulmonary hypertension] - (0.013*60) + 0.33 [NYHA=II] + 0.43 [abscess] + 0.44 [urgent surgery] + 0.64 [Staphylococcus spp.] = -0.57

$$P(mortality) = \frac{\exp(-0.57)}{1 + \exp(-0.57)} \approx 36\%$$

LVEF: left ventricular ejection fraction; NYHA. New York Hearth Assotiation

1

Table S8: Critical appraisal using PROBAST.

Domain	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11
Model information											
Author, Year and Model name	De Feo, 2012	Gaca, 2011, STSS score	Madeira, 2016	Gatti, 2017a, AEPEI original	Gatti, 2017a, AEPEI alternate	Gatti, 2017b, ANCLA	Martínez-Sellés, 2014, PALUSE	Olmos, 2017, RISK-E	Di Mauro, 2017, EndoSCORE	Fernández-Hidalgo, 2018, sp.ES-I	Fernández-Hidalgo, 2018, sp. ES-II
1. Participants											
Risk of Bias	**High**	**High**	**Unclear**	**Low**	**Low**	**Low**	**Low**	**Low**	**Unclear**	**Low**	**Low**
Applicability	**High**	**Low**	**Unclear**	**Low**	**Low**	**Low**	**Low**	**Low**	**Unclear**	**Low**	**Low**
1.1 Were appropriate data sources used?											
	PY	PY	PY	PY	PY	PY	PY	PY	PY	PY	PY
1.2 Were all inclusions and exclusions of participants appropriate?											
	N	PN	NI	PY	PY	PY	PY	PY	NI	PY	PY
Observations:	**De Feo, 2012:** The model was developed in a subgroup of patients. These participants represent a selected lower (or higher) risk sample of the original. **Gaca, 2011:** Excluded complete sites if data were missing in some variables, likely to have introduced bias but less important than excluding individual participants. **Madeira, 2016; Di Mauro, 2017:** No information about recruitment methods and exclusion criteria. **Gatti, 2017a:** The model was developed using only data from 2008 in France, because the data collection was particularly exhaustive, comprehensive, and accurate, we did not consider it could introduce bias.										
2. Predictors											
Risk of Bias	**Unclear**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**
Applicability	**Low**	**Low**	**Low**	**Unclear**	**Low**	**Low**	**Low**	**Low**	**Low**	**Unclear**	**Low**
2.1 Were predictors defined and assessed in a similar way for all participants?											
	Y	Y	Y	PY	PY	Y	Y	Y	Y	N	N
2.2 Were predictor assessments made without knowledge of outcome data?											
	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI
2.3 Are all predictors available at the time the model is intended to be used?											
	Y	Y	Y	PN	Y	Y	Y	Y	Y	Y	Y
Observations:	No author informed if predictor assessments was make without knowledge of outcome data, although we didn't penalized RoB if predictors assessed had an objective interpretation. **De Feo, 2012:** There were predictors assessed with subjective interpretation. **Gatti, 2017a (original); Fernández-Hidalgo. 2017 (ES-I):** Systolic pulmonary artery pressure predictor could be hard to recovery. **Fernández-Hidalgo. 2017:** Databases were not homogeneous, but authors did an effort to homogenize it, we did not penalized the RoB.										
3. Outcome											
Risk of Bias	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**
Applicability	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**
3.1 Was the outcome determined appropriately?											
	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3.2 Was a pre-specified or standard outcome definition used?											
	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3.3 Were predictors excluded from the outcome definition?											
	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3.4 Was the outcome defined and determined in a similar way for all participants?											

Domain	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11
Model information											
Author, Year and Model name	De Feo, 2012	Gaca, 2011, STSS score	Madeira, 2016	Gatti, 2017a, AEPEI original	Gatti, 2017a, AEPEI alternate	Gatti, 2017b, ANCLA	Martínez-Sellés, 2014, PALUSE	Olmos, 2017, RISK-E	Di Mauro, 2017, EndoSCORE	Fernández-Hidalgo, 2018, sp.ES-I	Fernández-Hidalgo, 2018, sp. ES-II
	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3.5 Was the outcome determined without knowledge of predictor information?											
	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI
3.6 Was the time interval between predictor assessment and outcome determination appropriate?											
	PY	PY	PY	PY	PY	PY	PY	PY	PY	PY	PY
Observations:	When the outcome is a hard variable which do not required interpretation such as mortality, previous knowledge of predictor information does not introduce RoB.										
4. Analysis											
Risk of Bias	**High**	**High**	**High**	**High**	**High**	**High**	**High**	**High**	**Unclear**	**Low**	**Low**
4.1 Were there a reasonable number of participants with the outcome?											
	N	Y	N	N	N	N	NI	N	PN	PN	PN
4.2 Were continuous and categorical predictors handled appropriately?											
	N	PY	PN	N	N	N	PN	N	PY	PY	PY
4.3 Were all enrolled participants included in the analysis?											
	PN	PN	PY	NI	PN	PN	PY	N	PY	PN	PN
4.4 Were participants with missing data handled appropriately?											
	NI	PN	NI	NI	NI	NI	NI	NI	NI	NI	NI
4.5 Was selection of predictors based on univariable analysis avoided?											
	N	N	N	N	N	N	N	N	N	Y	Y
4.6 Were complexities in the data accounted for appropriately?											
	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
4.7 Were relevant model performance measures evaluated appropriately?											
	N	PN	N	Y	PN	PN	N	Y	Y	Y	Y
4.8 Were model overfitting and optimism in model performance accounted for?											
	N	N	N	N	N	N	N	N	Y	Y	Y
4.9 Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?											
	PN	Y	Y	Y	Y	Y	PN	Y	Y	Y	Y
Observations:	**De Feo, 2012:** Very small number of events per parameter (EPP), continuous predictors not handled appropriately, probably using complete data and only apparent validation available. **Gaca, 2011:** Large EPP (aprox. 30), but predictors selected based on univariable analysis, random splitting validation (D:70% and V:30%) and no inform how missing data were handled. **Madeira, 2016:** Very small EPP and apparent validation. **Gatti 2017a; Gatti, 2017b:** Very small EPP, predictors selected based on univariable analysis and continuous predictors categorized based on the best discriminative performance. **Martínez-Sellés, 2014:** EPP not available, no informed about missing data, continuous predictors dichotomized and apparent performance. **Olmos, 2017:** Very small EPP and random splitting validation (D:70% and V:30%) and did not inform neither missing data nor continuous predictors were handled. **Di Mauro, 2017:** Predictors were selected based on univariable analysis ($p<0.2$). Although EPP was sufficiently large and model performance was optimism adjusted, unfortunately calibration measures were tested but not reported, thus we rated it as unclear RoB. **Fernández-Hidalgo, 2018:** EPP was slightly lower than required but were not univariable selection and model performance was optimism adjusted by bootstrap validation. The complete data analysis were not worrying because only 4 (0.5%) patients were excluded. We did not penalized RoB.										

Y: Yes; PY: Probably yes; N: No; PN: Probably no; NI: No information

Figure S1: Summary of risk of bias and applicability of the studies

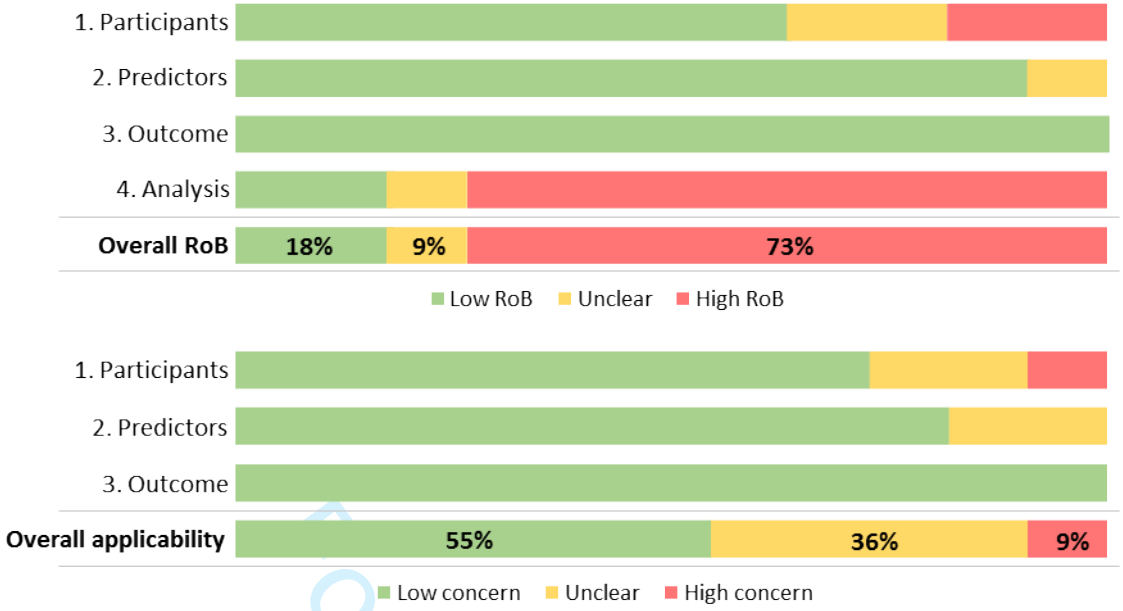
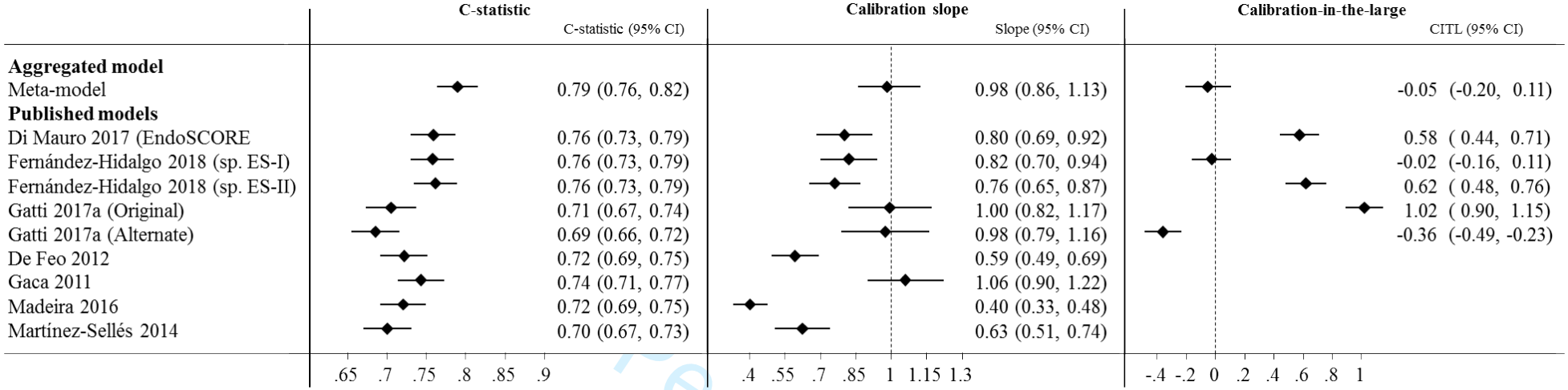
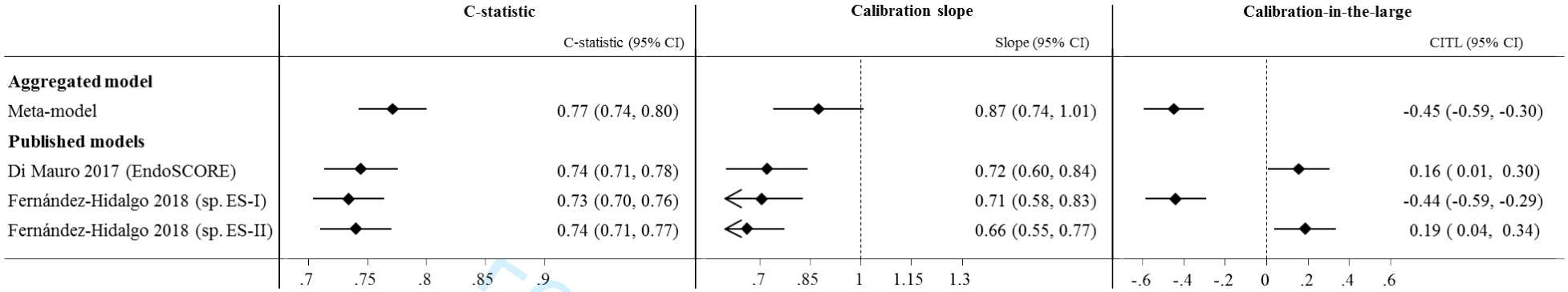


Figure S2: Validation of all models regardless of critical appraisal.



Dashed lines indicate lines of perfect calibration slope (1) and calibration-in-the-large (0). Black diamonds indicate point estimates and horizontal bars indicate 95% CIs.

Figure S3: Validation of the meta-model and existing models selected for aggregation for 30-days mortality outcome.



Dashed lines indicate lines of perfect calibration slope (1) and calibration-in-the-large (0). Black diamonds indicate point estimates and horizontal bars indicate 95% CIs.

S6: Members of GAMES group

Hospital Costa del Sol, (Marbella): Fernando Fernández Sánchez, Mariam Noureddine, Gabriel Rosas, Javier de la Torre Lima; **Hospital Universitario de Cruces**, (Bilbao): Elena Bereciartua, Roberto Blanco, María Victoria Boado, Marta Campaña Lázaro, Alejandro Crespo, Laura Guio Carrión, Mikel Del Álamo Martínez de Lagos, Gorane Euba Ugarte, Josune Goikoetxea, Marta Ibarrola Hierro, José Ramón Iruretagoyena, Josu Irurzun Zuazabal, Leire López-Soria, Miguel Montejo, Javier Nieto, David Rodrigo, Regino Rodríguez, Yolanda Vitoria, Roberto Voces; **Hospital Universitario Virgen de la Victoria**, (Málaga): Mª Victoria García López, Radka Ivanova Georgieva, Guillermo Ojeda, Isabel Rodríguez Bailón, Josefa Ruiz Morales; **Hospital Universitario Donostia-Poliklínica Gipuzkoa-IIS Biodonostia**, (San Sebastián): Harkaitz Azkune Galparsoro, Elisa Berritu Boronat, Mª Jesús Bustinduy Odriozola, Cristina del Bosque Martín, Tomás Echeverría, Alberto Eizaguirre Yarza, Ana Fuentes, Miguel Ángel Goenaga, Muskilda Goyeneche del Río, Ángela Granda Bauza, José Antonio Iribarren, Xabier Kortajarena Urkola, José Ignacio Pérez-Moreiras López, Ainhoa Rengel Jiménez, Karlos Reviejo, Alberto Sáez Berbejillo, Elou Sánchez Haza, Rosa Sebastián Alda, Itziar Solla Ruiz, Irati Unamuno Ugartemendia, Diego Vicente Anza, Iñaki Villanueva Benito, Mar Zabalo Arrieta; **Hospital General Universitario de Alicante**, (Alicante): Rafael Carrasco, Vicente Climent, Patricio Llamas, Esperanza Merino, Joaquín Plazas, Sergio Reus; **Complejo Hospitalario Universitario A Coruña**, (A Coruña): Nemesio Álvarez, José María Bravo-Ferrer, Laura Castelo, José Cuenca, Pedro Llinares, Enrique Miguez Rey, María Rodríguez Mayo, Efrén Sánchez, Dolores Sousa Regueiro; **Complejo Hospitalario Universitario de Huelva**, (Huelva): Francisco Javier Martínez; **Hospital Universitario de Canarias**, (Canarias): Mª del Mar Alonso, Beatriz Castro, Teresa Delgado Melian, Javier Fernández Sarabia, Dácil García Rosado, Julia González González, Juan Lacalzada, Lissete Lorenzo de la Peña, Alina Pérez Ramírez, Pablo Prada Arrondo, Fermín Rodríguez Moreno; **Hospital Regional Universitario de Málaga**, (Málaga): Antonio Plata Ciezar, José Mª Reguera Iglesias; **Hospital Universitario Central Asturias**, (Oviedo): Víctor Asensi Álvarez, Carlos Costas, Jesús de la Hera, Jonnathan Fernández Suárez, Lisardo Iglesias Fraile, Víctor León Arguero, José López Menéndez, Pilar Mencia Bajo, Carlos Morales, Alfonso Moreno Torrico, Carmen Palomo, Begoña Paya Martínez, Ángeles Rodríguez Esteban, Raquel Rodríguez García, Mauricio Telenti Asensio; **Hospital Clínic-IDIBAPS, Universidad de Barcelona**, (Barcelona): Manuel Almela, Juan Ambrosioni, Manuel Azqueta, Mercè Brunet, Marta Bodro, Ramón Cartañá, Carlos Falces, Guillermina Fita, David Fuster, Cristina García de la Mària, Delia García-Pares, Marta Hernández-Meneses, Jaume Llopis Pérez, Francesc Marco, José M. Miró, Asunción Moreno, David Nicolás, Salvador Ninot, Eduardo Quintana, Carlos Paré, Daniel Pereda, Juan M. Pericás, José L. Pomar, José Ramírez, Irene Rovira, Elena Sandoval, Marta Sitges, Dolors Soy, Adrián Téllez, José M. Tolosana, Bárbara Vidal, Jordi Vila; **Hospital General Universitario Gregorio Marañón**, (Madrid): Iván Adán, Juan Carlos Alonso, Ana Álvarez-Uría, Javier Bermejo, Emilio Bouza, Gregorio Cuerpo Caballero, Antonia Delgado Montero, Ana González Mansilla, Mª Eugenia García Leoni, Esther Gargallo, Víctor González Ramallo, Martha Kestler Hernández, Amaia Mari Hualde, Marina Machado, Mercedes Marín, Manuel Martínez-Sellés, Patricia Muñoz, María Olmedo, Álvaro Pedraz, Blanca Pinilla, Ángel Pinto, Cristina Rincón, Hugo Rodríguez-Abella, Marta Rodríguez-Créixems, Antonio Segado, Neera Toledo, Maricela Valerio, Pilar Vázquez, Eduardo Verde Moreno; **Hospital Universitario La Paz**, (Madrid): Isabel Antorrena, Belén Loeches, Mar Moreno, Ulises Ramírez, Verónica Rial Bastón, María Romero, Sandra Rosillo; **Hospital Universitario Marqués de Valdecilla**, (Santander): **Hospital Universitario Marqués de Valdecilla**, (Santander): Jesús Agüero Balbín, Cristina Amado, Carlos Armiñanzas Castillo, Ana Arnaiz, Francisco Arnaiz de las Revillas, Manuel Cobo Belaustegui, María Carmen Fariñas, Concepción Fariñas-Álvarez, Marta Fernández Sampedro, Iván García, Claudia González Rico, Laura Gutierrez-Fernandez , Manuel Gutiérrez-Cuadra, José Gutiérrez Díez, Marcos Pajarón, José Antonio Parra, Ramón Teira, Jesús Zarauza; **Hospital Universitario Puerta de Hierro**, (Madrid): Jorge Calderón Parra, Marta Cobo, Fernando Domínguez, Pablo García Pavía, Ana Fernández Cruz, Antonio Ramos-Martínez, Isabel Sánchez Romero; **Hospital Universitario Ramón y Cajal**, (Madrid): Tomasa Centella, José Manuel

Hermida, José Luis Moya, Pilar Martín-Dávila, Enrique Navas, Enrique Oliva, Alejandro del Río, Jorge Rodríguez-Roda Stuart, Soledad Ruiz; **Hospital Universitario Virgen de las Nieves**, (Granada): Carmen Hidalgo Tenorio; **Hospital Universitario Virgen Macarena**, (Sevilla): Manuel Almendro Delia, Omar Araji, José Miguel Barquero, Román Calvo Jambrina, Marina de Cueto, Juan Gálvez Acebal, Irene Méndez, Isabel Morales, Luis Eduardo López-Cortés; **Hospital Universitario Virgen del Rocío**, (Sevilla): Arístides de Alarcón, Encarnación Gutiérrez-Carretero, José Antonio Lepe, José López-Haldón, Rafael Luque-Márquez, Guillermo Marín, Antonio Ortiz-Carrellán, Eladio Sánchez-Domínguez; **Hospital San Pedro**, (Logroño): Luis Javier Alonso, Pedro Azcárate, José Manuel Azcona Gutiérrez, José Ramón Blanco, Antonio Cabrera Villegas, Lara García-Álvarez, Concepción García García, José Antonio Oteo; **Hospital de la Santa Creu i Sant Pau**, (Barcelona): Natividad de Benito, Mercé Gurguí, Cristina Pacho, Roser Pericas, Guillem Pons; **Complejo Hospitalario Universitario de Santiago de Compostela**, (A Coruña): M. Álvarez, A. L. Fernández, Amparo Martínez, A. Prieto, Benito Regueiro, E. Tijeira, Marino Vega; **Hospital Santiago Apóstol**, (Vitoria): Andrés Canut Blasco, José Cordo Mollar, Juan Carlos Gainzarain Arana, Oscar García Uriarte, Alejandro Martín López, Zuriñe Ortiz de Zárate, José Antonio Urturi Matos; **Hospital SAS Línea de la Concepción**, (Cádiz): Sánchez-Porto Antonio, Úbeda Iglesias Alejandro; **Hospital Clínico Universitario Virgen de la Arrixaca** (Murcia): José Mª Arribas Leal, Elisa García Vázquez, Alicia Hernández Torres, Ana Blázquez, Gonzalo de la Morena Valenzuela; **Hospital de Txagorritxu**, (Vitoria): Ángel Alonso, Javier Aramburu, Felicitas Elena Calvo, Anai Moreno Rodríguez, Paola Tarabini-Castellani; **Hospital Virgen de la Salud,** (Toledo): Eva Heredero Gálvez, Carolina Maicas Bellido, José Largo Pau, Mª Antonia Sepúlveda, Pilar Toledano Sierra, Sadaf Zafar Iqbal-Mirza; **Hospital Rafael Méndez,** (Lorca-Murcia):, Eva Cascales Alcolea, Ivan Keituqwa Yañez, Julián Navarro Martínez, Ana Peláez Ballesta; **Hospital Universitario San Cecilio** (Granada): Eduardo Moreno Escobar, Alejandro Peña Monje, Valme Sánchez Cabrera, David Vinuesa García; **Hospital Son Llátzer** (Palma de Mallorca): María Arrizabalaga Asenjo, Carmen Cifuentes Luna, Juana Núñez Morcillo, Mª Cruz Pérez Seco, Aroa Villoslada Gelabert; **Hospital Universitario Miguel Servet** (Zaragoza): Carmen Aured Guallar, Nuria Fernández Abad, Pilar García Mangas, Marta Matamala Adell, Mª Pilar Palacián Ruiz, Juan Carlos Porres; **Hospital General Universitario Santa Lucía** (Cartagena): Begoña Alcaraz Vidal, Nazaret Cobos Trigueros, María Jesús Del Amor Espín, José Antonio Giner Caro, Roberto Jiménez Sánchez, Amaya Jimeno Almazán, Alejandro Ortín Freire, Monserrat Viqueira González; **Hospital Universitario Son Espases** (Palma de Mallorca): Pere Pericás Ramis, Mª Ángels Ribas Blanco, Enrique Ruiz de Gopegui Bordes, Laura Vidal Bonet; **Complejo Hospitalario Universitario de Albacete** (Albacete): Mª Carmen Bellón Munera, Elena Escribano Garaizabal, Antonia Tercero Martínez, Juan Carlos Segura Luque; **Hospital Universitario Terrassa:** Cristina Badía, Lucía Boix Palop, Mariona Xercavins, Sónia Ibars. **Hospital Universitario Dr. Negrín** (Gran Canaria): Xerach Bosch, Eloy Gómez Nebreda, Ibalia Horcajada Herrera, Irene Menduiña Gallego, Imanol Pulido; **Complejo Hospitalario Universitario Insular Materno Infantil** (Las Palmas de Gran Canaria): Héctor Marrero Santiago, Isabel de Miguel Martínez, Elena Pisos Álamo. **Hospital Universitario 12 de Octubre** (Madrid): Eva Mª Aguilar Blanco, Mercedes Catalán González, María Angélica Corres Peiretti, Andrea Eixerés Esteve, Laura Domínguez Pérez, Santiago de Cossío Tejido, Francisco Galván Román, José Antonio García Robles, Francisco López Medrano, Mª Jesús López Gude, Mª Ángeles Orellana Miguel, Patrick Pilkington, Yolanda Revilla Ostalaza, Juan Ruiz Morales, Sebastián Ruiz Solís, Ana Sabín Collado, Marcos Sánchez Fernández, Javier Solera Rallo, Jorge Solís Martín. **Hospital Universitari de Bellvitge (L'Hospitalet de Llobregat):** Guillermo Cuervo, Francesc Escrihuela-Vidal, Jordi Carratalà, Inmaculada Grau, Sara Grillo, Carmen Ardanuy, Dámaris Berbel, José Carlos Sánchez Salado, Oriol Alegre, Alejandro Ruiz Majoral, Fabrizio Sbraga, Arnau Blasco, Laura Gracia Sánchez, Iván Sánchez-Rodríguez. **Hospital Universitario Fundación Jiménez Díaz** (Madrid): Beatriz Álvarez, Alfonso Cabello Úbeda, Ricardo Fernández Roblas, Miguel Ángel Navas Lobato, Ana María Pello. **Hospital Basurto** (Bilbao): Mireia de la Peña Triguero, Ruth Esther Figueroa Cerón, Lara Ruiz Gómez. **Hospital del Mar** (Barcelona): Mireia Ble, Juan Pablo Horcajada Gallego, Antonio José Ginel, Inmaculada López, Alexandra Mas, Antoni Mestres, Lluís

Infection
https://doi.org/10.1007/s15010-019-01338-x

REVIEW

Prognostic factors of mortality after surgery in infective endocarditis: systematic review and meta-analysis

Laura Varela Barca[1] · Enrique Navas Elorza[2] · Nuria Fernéndez-Hidalgo[3] · Jose Luis Moya Mur[4] · Alfonso Muriel García[5] · B. M. Fernández-Felix[5,6] · Javier Miguelena Hycka[1] · Jorge Rodríguez-Roda[1] · Jose López-Menéndez[1]

Received: 25 March 2019 / Accepted: 22 June 2019

Abstract

Purpose There is a lack of consensus about which endocarditis-specific preoperative characteristics have an actual impact over postoperative mortality. Our objective was the identification and quantification of these factors.

Methods We performed a systematic review of all the studies which reported factors related to in-hospital mortality after surgery for acute infective endocarditis, conducted according to PRISMA recommendations. A search string was constructed and applied on three different databases. Two investigators independently reviewed the retrieved references. Quality assessment was performed for identification of potential biases. All the variables that were included in at least two validated risk scores were meta-analyzed independently, and the pooled estimates were expressed as odds ratios (OR) with their confidence intervals (CI).

Results The final sample consisted on 16 studies, comprising a total of 7484 patients. The overall pooled OR were statistically significant ($p < 0.05$) for: age (OR 1.03, 95% CI 1.00–1.05), female sex (OR 1.56, 95% CI 1.35–1.81), urgent or emergency surgery (OR 2.39 95% CI 1.91–3.00), previous cardiac surgery (OR 2.19, 95% CI 1.84–2.61), NYHA ≥ III (OR 1.84, 95% CI 1.33–2.55), cardiogenic shock (OR 4.15, 95% CI 3.06–5.64), prosthetic valve (OR 1.98, 95% CI 1.68–2.33), multivalvular affection (OR 1.35, 95% CI 1.01–1.82), renal failure (OR 2.57, 95% CI 2.15–3.06), paravalvular abscess (OR 2.39, 95% CI 1.77–3.22) and *S. aureus* infection (OR 2.27, 95% CI 1.89–2.73).

Conclusions After a systematic review, we identified 11 preoperative factors related to an increased postoperative mortality. The meta-analysis of each of these factors showed a significant association with an increased in-hospital mortality after surgery for active infective endocarditis.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s15010-019-01338-x) contains supplementary material, which is available to authorized users.

✉ Laura Varela Barca
lauravarela21089@gmail.com

[1] Department of Cardiovascular Surgery, University Hospital Ramon y Cajal, Ctra. Colmenar Viejo, km. 9.100, 28034 Madrid, Spain

[2] Department of Infectology, University Hospital Ramon y Cajal, Madrid, Spain

[3] Department of Infectious Diseases, University Hospital Vall d'Hebron, Barcelona, Spain

[4] Department of Cardiology, University Hospital Ramon y Cajal, Madrid, Spain

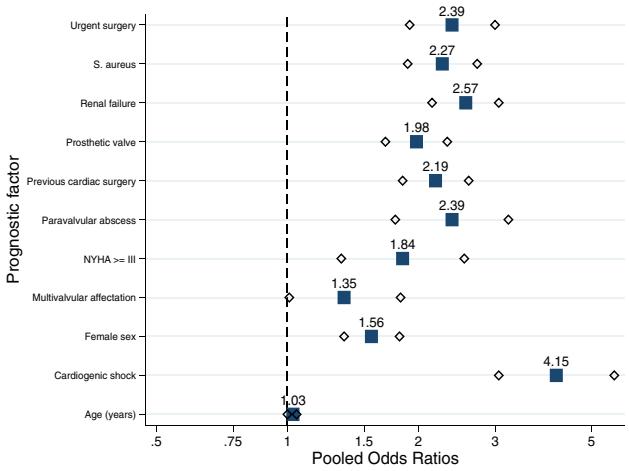
[5] Clinical Biostatistics Unit, Hospital Ramon y Cajal (IRYCIS), Madrid, Spain

[6] CIBER Epidemiology and Public Health (CIBERESP), Madrid, Spain

Published online: 28 June 2019

Graphic abstract

Graph summary of the Pooled Odds Ratios of the 11 preoperative factors analyzed after the systematic review and meta-analysis.



Keywords Infective endocarditis · Prognostic factors · Systematic review · Meta-analysis

Introduction

The mortality rate in patients operated on after the diagnosis of infective endocarditis (IE) is reported to range from 15% to more than 45%, depending on several factors, such as patient baseline characteristics, preoperative status, offending pathogen, intraoperative difficulties and hospital expertise variations [1]. In spite of the improvement of antimicrobial therapy, the advances in surgical techniques and the shortening of waiting time to surgery, the mortality associated with IE continues to be very high [1, 2]. In that context, although cardiac surgery is an essential procedure in the treatment of IE, a proportion of patients who have surgical indication do not undergo surgical treatment because of their high surgical risk, which could lead to surgery rejection [3–6]. It is estimated that less than half of the patients who have a surgical indication finally undergo surgical intervention [7–10]. However, it is well known that patients who have indication for surgery and are not operated on have a dismal prognosis [11], whereas long-term survival in patients who survive the cardiac surgery is acceptable [12].

Therefore, accurate surgical risk estimation is crucial for the surgical decision-making process. In the recent years, several new IE-specific risk scores have been developed. They incorporate some IE-specific factors that are thought to be independent mortality factors [8, 13]. However, the impact of all these specific factors in postoperative mortality continues to be in doubt.

The aim of the present systematic review and meta-analysis was to clarify which are the main pre-operative factors related to in-hospital mortality after valve surgery for active IE. An adequate knowledge of the impact on survival of these factors may help to guide the surgical decision process, through the identification of patients with a higher risk of suffering a poor outcome after surgery.

Methods

Data sources

The systematic review was undertaken in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [14], and it followed the recommendations of Meta-analysis Of Observational Studies in Epidemiology (MOOSE) Group [15]. The literature search was carried out with the help of an experienced librarian. The review question and the complete search strategy [16] are detailed in Fig. 1-1.

Study selection

The inclusion criteria for the eligible studies were studies whose study population comprised adult subjects with active IE who underwent valve surgery. The investigation of the association between a prognostic factor and in-hospital postoperative mortality was required for inclusion, or raw data from which this association could be determined instead. Active endocarditis was defined as on-going active infection under antimicrobial treatment at the time of surgery [17].

Review question (PICOTS):

Patients: Adult subjects (≥18 years) with active IE who underwent surgical treatment.
Index models/studies: Prediction models with and without external validation and studies that analyzed factors related with in-hospital mortality.
Comparator: n/a
Outcomes: In-hospital mortality (or 30-day mortality)
Timing: From surgery to 30 days in the post-operatory period or until patient discharge.
Setting: Factors that preoperatively have influence on mortality after surgery.

Complete search strategy: (“endocarditis” OR "Endocarditis"[Mesh]) AND ("Cardiac Surgical Procedures"[Mesh] OR “cardiac surgery”) AND ("Prognosis"[Mesh] OR “prognostic factor”) AND ("Mortality"[Mesh] OR “mortality”) AND ("01/01/2000"[Date - Publication] : "3000"[Date - Publication]) AND ("english"[Language] OR "spanish"[Language]) AND ("clinical study"[Publication Type] OR "comparative study"[Publication Type] OR "observational study"[Publication Type] OR "multicenter study"[Publication Type] OR "clinical trial"[Publication Type] OR "meta-analysis"[Publication Type]).

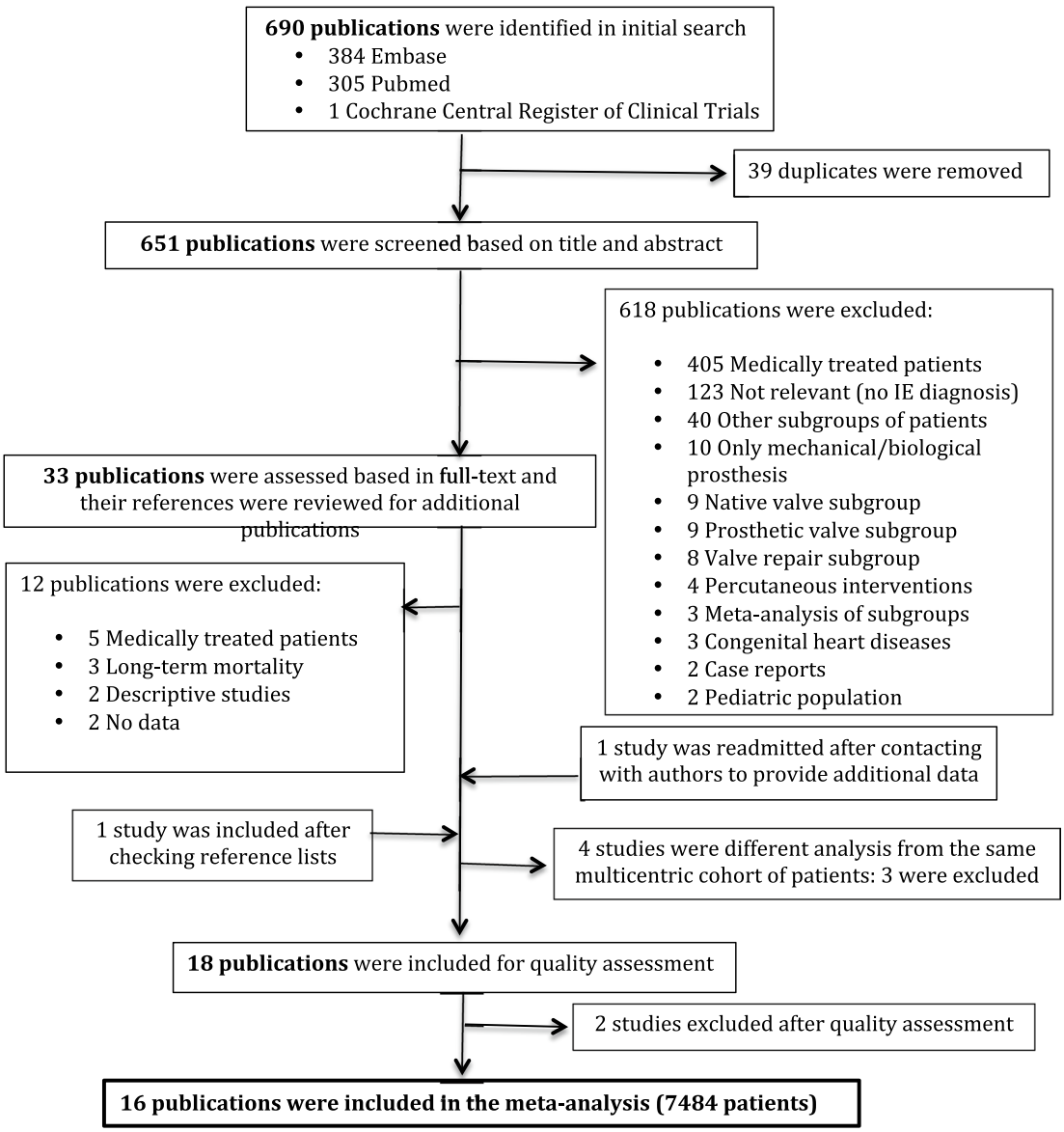


Fig. 1 Search strategy. 1: Review question and complete search strategy. 2: Flow chart of included/excluded studies

Observational studies, randomized controlled trials, and systematic reviews and meta-analysis were included. Redundancies and overlaps between studies were managed and controlled after quality assessment. Studies that included participants who were managed with medical treatment alone, and those studies focused exclusively on specific subgroups (e.g., valve repair, native valves, congenital diseases, etc.) were excluded. The main reason for exclusion of those specific subgroups publications was the lack of availability of full-sample data, with a high risk of selection bias. The search was limited to articles published in English or Spanish languages, in peer-reviewed journals, involving human subjects. In addition, the bibliographic search was limited to studies published from 2000, because of the change in surgical indications and diagnosis with the development of the IE guidelines [18–20]. The selected search string was applied on three databases (Medline, Cochrane and EMBASE). Meeting abstracts, case reports, conference presentations, editorials, and expert opinions were excluded.

Figure 1-2 shows the flow chart of the process of selection of the eligible references, and the reasons for inclusion/exclusion at each step are summarized.

Data extraction and quality assessment

Two reviewers (JLM and LVB) independently reviewed the full-text articles of the remaining references, and eligibility against the predefined criteria was evaluated. Discrepancies between the two reviewers were resolved by discussion and consensus. Complete References list of the eligible studies was checked for possible additional unidentified references. Finally, information was extracted from article text, tables, and figures of each selected study (Table 2).

The methodological quality of the selected studies was assessed using the Quality in Prognostic Studies tool (QuIPS), using the criteria previously published by Hayden et al. [21]. Studies with an unacceptable risk of bias were excluded, considered as those with more than one aspect classified as "high risk of bias".

Prognostic factor selection and definitions

Through the systematic review of the literature, we identified the most common IE-specific variables (Table 1). All those variables that were considered as risk factors for mortality in two or more IE-specific scores were included in the meta-analysis. Therefore, a total of 11 possible pre-operative prognostic factors related to pre-surgical conditions were identified:

1. Age (considered as a continuous variable, expressed in years).
2. Female sex.
3. Urgent surgery: surgery required within 24 h of its indication [22].
4. Emergency surgery: surgery required on the day of admission [22].
5. Previous cardiac surgery: previous surgical procedure with opening of the pericardium.
6. Functional class ≥ III before valve surgery, according to New York Heart Association (NYHA) classification [23].
7. Cardiogenic shock: acute myocardial dysfunction, with systolic pressure < 90 mmHg, tissue hypoperfusion and low cardiac output [24].
8. Prosthetic valve IE: IE affecting a previous prosthetic valve.
9. Multivalvular involvement: IE affecting more than one heart valve.
10. Renal failure: presence of a serum creatinine concentration > 2 mg/dl before surgery.
11. Paravalvular abscess: purulent cavity with necrosis and capacity to invade adjacent structures [25].
12. *Staphylococcus aureus* as the causative agent of the IE episode.

The analyzed outcome was postoperative death, considered as in-hospital mortality and/or death in the first 30 days after surgery [26, 27].

Statistical analysis

All statistical analyses were performed using Stata/IC 14.2 (Stata Statistical Software: Release 14. College Station, TX: StataCorp LP), implemented with the meta-analysis OR, RR, RD, IR, ID, B, MD & R Combined: User-written command [28].

The meta-analysis was performed pooling all the data of reported in hospital mortality in every study, stratified by each of the previously identified IE-specific prognostic factors.

The calculated univariate OR for in-hospital mortality of each of the selected studies were used as the individual summary statistic to obtain the pooled estimation. Heterogeneity across the studies was assessed using the I^2 statistic. An I^2 greater than 25% was considered as substantial heterogeneity [29, 30]. The weighting method used was the inverse of variance fixed-effects model (FEM), if there was no significant heterogeneity across the studies, or random-effects model (REM) if substantial heterogeneity was observed [31].

Forest plot graphs were used to graphically depict the association between early mortality and each of the analyzed factors. Empirical correction of zeros methodology was used if zero mortality events were observed in 1 group. However, forest plot graphs were presented after excluding studies with zero mortality for graphical quality improvement.

Table 1 Variables included in the IE-specific scores (STS-IE, PALSUSE, De Feo-Cotrufo, Costa score, Risk-E, AEPEI, Endoscore and Specific Euroscores I and II) with the corresponding value given in each of them

	STS-IE	PALSUSE	De Feo-Cotrufo	Costa	Risk-E	AEPEI	Endoscore	SpecificES1'	SpecificES2'
Age		> 70:1	40–49:5 50–59:7 60–69:9 70–79:11 > 80:13	≥ 40:4	≤ 51:0 61:9 65:13 ≥ 73:14		60–70:0.46 70–80:0.88 > 80:1.53	≤ 60:0 65:11 70:22 75:33 80:44 90:67 100:78	≤ 60:0 65:11 70:22 75:33 80:44 90:67 100:78
Sex (female)		1					0.51		
Urgent surgery	6	1						Emergent:30	42 Emergent:70
Cardiogenic shock	17 In/IABP:10	∂	Critical state:11 NYHA IV:9	5	15	Critical:1.5 NYHA IV:1.3	1.46	42	37
Multival-vular	9						2:0.5 3:1.5		
Prosthetic IE		1	Not appli-cable		6				
Prior cardiac surgery	CABG:7 Valve:7	∂						42	35
Renal failure	12	∂	5		5	2.2	0.5	18	
Arrhythmia	8			8 AV block:5					
Active IE	10								
Positive blood cultures		*S. aureus*:1	5		*Virulent*:9		*P. auregi-nosa:*1.46 *S. aureus*:1.24 Fungi:1.66 Other:0.60	*Staphyl*:20	*Staphyl*:19
Abscess cardiac destruct		1	5	TEE anom:5	5		1.09	Fistulae:46	Fistulae:41
Other vari-able	IDDM:8 NIDDM:6 Hypert:5 COPD:5	ES ≥ 10%:1		Veg ≥ 10:4 Sepsis:6	Trombop:7 Septic shock:7	BMI > 27:1 sPAP > 55:1	LVEF:-0.03	NYHA > 1:26 Pulm HT:48 Mitral:14	NYHA > 1:22 Mitral:14

Fernández–Hidalgo et al. study

IE infective endocarditis, *In* inotropes, *IABP* intra-aortic balloon pump, *CABG* coronary artery bypass grafting, *Hyper* high blood pressure, *COPD* chronic obstructive pulmonary disease, *ES* EuroScore, *NYHA* New York Heart Association, *AV* atrio–ventricular, *TEE* transesophageal echocardiography, *Veg* vegetations, *Trombop* thrombocytopenia, *BMI* body mass index, *sPAP* systolic pulmonary artery pressure, *LVEF* left ventricular ejection fraction, *Pulm HT* pulmonary hypertension

[∂]Included in EuroSCORE calculation

Publication bias was assessed using Egger method [32] and Funnel plot graphs [33]. All *p* values were two-sided. In addition, to assess the influence of each individual study on the pooled estimates, we performed an influence analysis by sequentially removing each individual study, to analyze the robustness of the estimated ORs.

Results

Description of included studies and patients

Figure 1-2 shows the flow chart for study selection. After full-text review, 18 observational studies were included [9,

Table 2 Main characteristics of included studies

Study	Pub year	Design	Outcomes	Prognostic factors	Sample size	IE	Age	Mort.	Native prosthetic ID	Valve affected
David et al.	2007	RU	In-hospital[m] Long term[m]	3, 6–11	383	LR	51 ± 16	12.0% (45)	69%N 31%P	56%A 27%M 15%MA
Mestres et al.	2007	RU	In-hospital[m] EuroSCORE validation	1–5, 8, 11	180	LR	57.2 ± 15.6	28.8% (55)	n/a	57.1%A 25.7%M 9.9%MA
Hanai et al.	2008	RU	In-hospital[m]	1, 2, 4, 5, 7, 8, 10, 11	94	LR	50.1 ± 15.9	15.0% (14)	84.04%N 17.02%P	55.32%A 29.78%M 11.70%MA
Fayad G. et al.	2011	RU	In-hospital[m]	1, 2, 5, 7, 8, 10, 11	141	LR	56.3 ± 14.9	16.0% (22)	87%N 13%P	39%A 34.75%M 9.93%MA
Caes et al.	2013	RU	In-hospital[m] Long term[m]	1–5, 9, 11	186	LR	57.8 ± 15.3	13.0% (24)	82%N 18%P	45%A 35%M 16%MA
Hussain et al.	2014	RU	In-hospital[m] Long term[m]	1, 7, 10	775	L	n/a	8.0% (62)	53%N 47%P	51%A 31%M 18%MA
Martínez-Sellés et al.	2014	PM	In-hospital[m] Score develop.	1–3, 7, 10, 11	437	LR ID	65.9 ± 16.4	24.3% (106)	61,1%N 27,9%P 11%ID	52.6% A 42.1%M n/a
Spilopoulos et al.	2014	RU	In-hospital[m] Long term[m]	1–3, 5, 7, 8, 9–11	94	LR	58.3 ± 13.1	8.5% (8)	90.4%N 9.6%P	57.4%A 29.8%M 8,5%MA
Marks et al.	2014	RU	30-days[m]	1, 2, 4	336	LR	52	12.2% (41)	79.2%N 20.8%P	56%A 53.9%M 23.5%MA
Oh et al.	2014	RU	In-hospital[m] Long term[m] Morbility Scores valid	1, 2, 4–6, 8–11	146	LR	48.8 ± 16	6.8% (10)	66,4%N 33.6%P 14.4%ID	64.4%A 42.5%M 14.4%MA
Olmos et al.	2017	PM	In-hospital[m] Score develop.	1, 2, 4–6, 8–11	424	L	61	29.3% (124)	61.1%N 38.9%P	n/a
Di Mauro et al.	2017	RM	In-hospital[m]	1, 2, 4–6, 7–11	2715	LR	59.6 ± 15.1	11.0% (298)	79.6%N 20.4%P	42.5%A 33%M 18%MA
Farag et al.	2017	RU	30 days [m]	1–5, 7–10	360	LR	58.7 ± 14.7	18.3% (68)	86.1%N 13.9%P	45.6%A 31.1%M 13.3%MA
Perrota et al.	2017	RU	In-hospital[m] Medium term[m] Morbility	1, 2, 7, 8, 10	254	LR	56 ± 16	8.9% (22)	72%N 28%P	57%A 31%M 9%MA

Table 2 (continued)

Study	Pub year	Design	Outcomes	Prognostic factors	Sample size	IE	Age	Mort.	Native prosthetic ID	Valve affected
Varela et al.	2017	RU	In-hospital[m] Scores valid	1–3, 5–11	180	L	61.8 ± 14.3	26.8% (48)	62.78%N 37.22%P	36.84%A 36.84%M 26.32%MA
Fernández-Hidalgo et al.	2018	RM	30 days [m] Score develop	1–6, 8–11	779	LR	58 ± 15.4	26.8% (208)	n/a	62.5%A 41.8%M 12.2%MA

Age: years ± standard deviation

Pub Year publication year, *IE* infective endocarditis, *R* retrospective, *P* prospective, *U* unicentric, *M* multicentric, *L* left, *R* right, *ID* implantable devices, [m] mortality, *PO* post-operative, *A* aortic, *M* mitral, *MA* mitro-aortic, *Mort.* mortality (% (number of patients))

^ 1: age, 2: female sex, 3: urgent surgery, 4: previous surgery, 5:NYHA ≥ III, 6: cardiogenic shock, 7: prosthetic valve, 8: multivalvular, 9: renal failure, 10:abscess, 11:*S. aureus*

10, 34–49]. The demographic details, distribution, methods and designs of these studies are shown in Table 2.

Four of the initially selected studies analyzed the same multipurpose database, in three different time periods [2, 9, 50, 51]. Therefore, to avoid selection bias because of repeated studies from the same cohort, only one study of these series was included [9].

Quality assessment

To assess the potential biases in the included studies, QuIPS [21] was employed by two reviewers independently in the selected references (Fig. 2-1). Two studies were excluded from the meta-analysis due to a moderate to severe risk of bias [34, 38]. Rob summary table [52] adapted to observational studies was used to show overall risk of bias (Fig. 2-2).

Outcomes

1. Age

   Age was analyzed in 15 of the included studies; however, the pooled estimation was only possible in five of them, where age was considered as a continuous variable, comprising a total of 1393 patients. Significant heterogeneity was found across the studies (I^2 53.10%; $p = 0.07$). The overall REM pooled OR (Fig. 3-1) was statistically significant (OR 1.03, 95% CI 1.00–1.05, $p < 0.01$).
2. Female sex

   Fourteen studies reported enough data to estimate the association of sex with the event of interest, comprising a total of 6326 patients (Fig. 3-2). I^2 was 3.96% ($p = 0.41$). The overall FEM pooled OR was 1.56 (95% CI 1.35–1.81, $p < 0.001$).
3. Urgent or emergency surgery

   9 studies analyzed the relation between urgent/emergent surgery and in-hospital mortality (2809 patients). I^2 resulted 38.27% ($p = 0.12$). The overall pooled OR after zero correction was statistically significant (OR 2.39, 95% CI 1.91–3.00, $p < 0.001$) (Fig. 3-3).
4. Previous cardiac surgery

   It was analyzed in 8 of the included studies (4796 patients). I^2 was 24.90%, ($p = 0.23$). The overall pooled OR was 2.19 (95% CI 1.84–2.61, $p < 0.001$) (Fig. 3-4).
5. NHYA functional class ≥ III

   NHYA ≥ III was considered in 10 studies (5119 patients). I^2 resulted 32.90% ($p = 0.15$). The REM pooled OR, adjusted for zero correction, was 1.84 (95% CI 1.33–2.55, $p = 0.002$) (Fig. 3-5).
6. Cardiogenic shock

   Acute circulatory failure was analyzed in 6 studies, comprising a total of 4627 patients. There was significant heterogeneity across the studies ($I^2 = 43.82\%$,

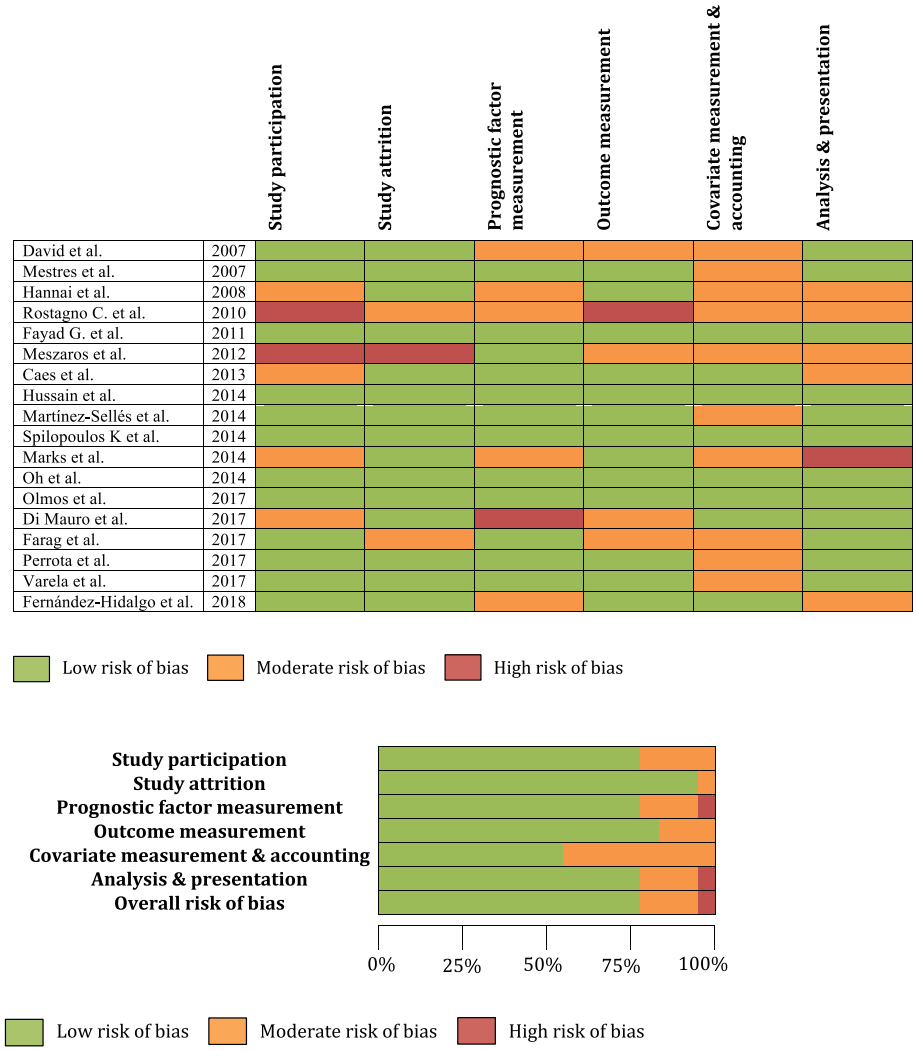


Fig. 2 Quality assessment. 1: Risk of bias table and 2: Rob summary

$p = 0.11$). REM pooled estimation resulted in a significant association (OR 4.15, 95% CI 3.06–5.64, $p < 0.001$) (Fig. 3-6).

7. Prosthetic valve IE

   Among the 16 studies included in the analysis, 11 analyzed prosthetic valve IE comprising a total of 5857 patients. No heterogeneity was observed across the studies (I^2 0%, $p = 0.89$). The overall pooled OR was 1.98 (95% CI 1.68–2.33, $p < 0.001$) (Fig. 3-7).

8. Multivalvular involvement

   Multivalvular IE was analyzed in 12 studies (5750 patients). The meta-analysis showed a significant heterogeneity (I^2 37.92%, $p = 0.11$). REM pooled estimation (after zero correction) showed a significant association (OR 1.35, 95% CI 1.01–1.82, $p = 0.003$) (Fig. 3-8).

9. Renal failure

   9 studies analyzed the relation between renal failure and in-hospital mortality, comprising a total of 5267 patients. The overall pooled OR was statistically significant (OR 2.57, 95% CI 2.15–3.06, $p < 0.001$) without heterogeneity across the studies (I^2 0%, $p = 0.87$) (Fig. 3-9).

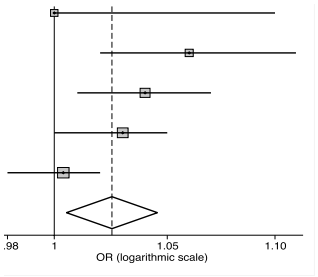
10. Paravalvular abscess

    Figure 3-10 shows the estimated OR of in-hospital mortality and abscess formation. Twelve analyzed paravalvular abscess, comprising a total of 6422 patients. The pooled OR was statistically significant (OR 2.39, 95% CI 1.77–3.22, $p < 0.001$) using REM (I^2 64.68%, $p = 0.001$).

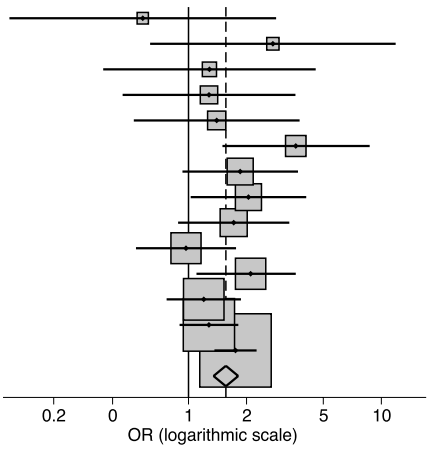
11. *Staphylococcus aureus* infection

    Staphylococcus aureus as the causative agent of the IE episode was analyzed in 11 studies (5759 patients). No heterogeneity was observed across the studies (I^2 0%, $p = 0.48$). The overall pooled OR was statistically significant (OR 2.27, 95% CI 1.89–2.73, $p < 0.001$) (Fig. 3-11).

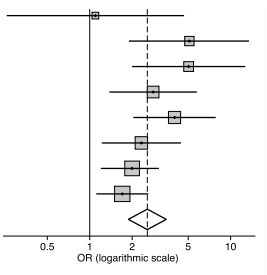
Study	Year	OR	95% CI Lower - Upper	Weight (%)
Caes et al.	2013	1.02	1.00 – 1.05	11.9
Perrota et al.	2017	1.01	1.00 – 1.04	14.0
Varela et al.	2017	1.02	0.99 – 1.05	21.4
Martínez-Sellés et al.	2014	1.02	0.99 – 1.05	24.6
Marks et al.	2014	1.03	1.01 – 1.05	28.1
Total		**1.03**	**1.00 – 1.05**	**100**



Study	Year	OR	95% CI Lower - Upper	Weight (%)
Oh et al.	2014	0.58	0.11 – 2.85	0.86
Spilopoulos K et al.	2014	2.74	0.63 – 11.87	1.01
Hanai et al.	2008	1.28	0.36 – 4.56	1.35
Fayad G. et al.	2011	1.28	0.46 – 3.58	2.05
Perrota et al.	2017	1.4	0.52 – 3.77	2.23
Caes et al.	2013	3.6	1.5 – 8.7	2.82
Marks et al.	2014	1.85	0.93 – 3.69	4.58
Varela et al.	2017	2.05	1.02 – 4.08	4.59
Mestres et al.	2007	1.72	0.88 – 3.33	4.94
Farag et al.	2017	0.97	0.53 – 1.76	6.11
Martínez-Sellés et al.	2014	2.1	1.1 – 3.6	6.20
Olmos et al.	2017	1.2	0.77 – 1.87	11.1
Fernández-Hidalgo et al.	2018	1.28	0.89 – 1.81	17.6
Di Mauro et al.	2017	1.75	1.36 – 2.25	34.4
Total		**1.56**	**1.35 – 1.81**	**100**



Study	Year	OR	95% CI Lower - Upper	Weights
Spilopoulos K et al.	2014	1.09	0.26 – 4.67	4.06
Caes et al.	2013	5.1	1.9 – 13.5	7.79
Fernández-Hidalgo et al.	2018	5.04	1.99 – 12.69	8.51
Varela et al.	2017	2.81	1.38 – 5.76	12.3
Mestres et al.	2007	4	2.04 – 7.84	13.2
David et al.	2007	2.32	1.21 – 4.44	13.9
Martínez-Sellés et al.	2014	2	1.2 – 3.1	19.2
Olmos et al.	2017	1.70	1.11 – 2.60	21.1
Total		**2.56**	**1.87 – 3.49**	**100**



Study	Year	OR	95% CI Lower - Upper	Weights
Oh et al.	2014	2.48	0.68 – 9.07	1.81
Hanai et al.	2008	2.62	0.76 – 9.01	1.98
Marks et al.	2014	3.85	1.15 – 12.5	2.13
Caes et al.	2013	2.4	1 – 6.1	3.71
Mestres et al.	2007	2.23	1.16 – 4.31	7.02
Farag et al.	2017	1.36	0.77 – 2.41	9.28
Fernández-Hidalgo et al.	2018	3.11	2.22 – 4.35	26.8
Di Mauro et al.	2017	1.87	1.45 – 2.42	47.3
Total		**2.19**	**1.84 – 2.61**	**100**

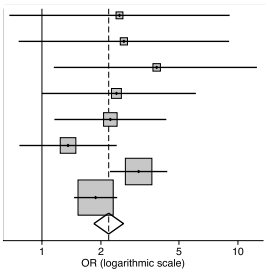
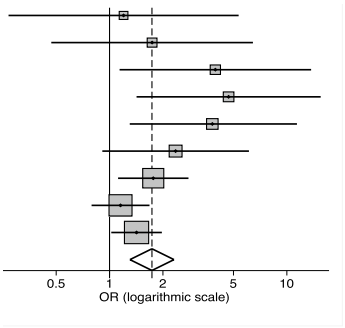
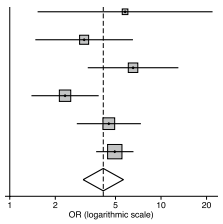


Fig. 3 Forest plot graphs. 1: Age, 2: Female sex, 3: Urgent surgery, 4: Previous cardiac surgery, 5: NYHA ≥ III, 6: Cardiogenic shock, 7: Prosthetic valve IE, 8: Multivalvular IE, 9: Renal failure, 10: Paravalvular abscess, 11: *S. aureus*. **OR* odds ratio; *CI* confidence interval

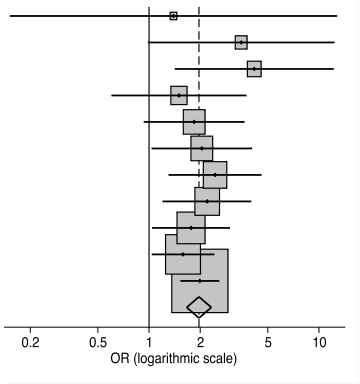
Study	Year	OR	95% CI Lower - Upper	Weights
Spilopoulos K et al.	2014	1.2	0.27 – 5.35	3.36
Oh et al.	2014	1.74	0.47 – 6.44	4.26
Hanai et al.	2008	3.95	1.14 – 13.71	4.67
Farag et al.	2017	4.69	1.42 – 15.52	5.01
Caes et al.	2013	3.8	1.3 – 11.4	5.91
Fayad G. et al.	2011	2.36	0.91 – 6.11	7.36
Olmos et al.	2017	1.76	1.11 – 2.79	19.7
Di Mauro et al.	2017	1.15	0.79 – 1.68	23.6
Fernández-Hidalgo et al.	2018	1.42	1.02 – 1.97	26.1
Total		**1.73**	**1.30 – 2.30**	**100**



Study	Year	OR	95% CI Lower - Upper	Weights
Oh et al.	2014	5.78	1.52 – 21.92	4.70
Varela et al.	2017	3.10	1.48 – 6.53	12.1
David et al.	2007	6.53	3.28 – 13.02	13.4
Olmos et al.	2017	2.32	1.39 – 3.82	19.2
Fernández-Hidalgo et al.	2018	4.52	2.77 – 7.35	20.2
Di Mauro et al.	2017	4.95	3.72 – 6.57	30.5
Total		**4.15**	**3.06 – 5.64**	**100**



Study	Year	OR	95% CI Lower - Upper	Weights
Spilopoulos K et al.	2014	1.39	0.15 – 12.79	0.56
Hanai et al.	2008	3.48	0.98 – 12.35	1.74
Fayad G. et al.	2011	4.16	1.42 – 12.22	2.40
Perrota et al.	2017	1.5	0.60 – 3.75	3.32
Farag et al.	2017	1.84	0.93 – 3.65	5.96
Varela et al.	2017	2.05	1.04 – 4.04	6.00
David et al.	2007	2.25	1.30 – 4.59	6.99
Martínez-Sellés et al.	2014	2.2	1.2 – 4	7.68
Hussain et al.	2014	1.77	1.04 – 2.99	9.97
Olmos et al.	2017	1.59	1.04 – 2.42	15.4
Di Mauro et al.	2017	1.99	1.52 – 2.59	39.9
Total		**1.98**	**1.68 – 2.33**	**100**



Study	Year	OR	95% CI Lower - Upper	Weights
Hanai et al.	2008	0.54	0.06 – 4.57	1.81
Farag et al.	2017	0.71	0.08 – 6.00	1.81
Fayad G. et al.	2011	1.55	0.39 – 6.08	4.08
Perrota et al.	2017	1.67	0.46 – 6.14	4.44
Mestres et al.	2007	1.15	0.41 – 3.22	6.62
Varela et al.	2017	4.42	2.12 – 9.20	10.8
David et al.	2007	1.33	0.67 – 2.67	11.6
Olmos et al.	2017	0.88	0.53 – 1.44	16.8
Fernández-Hidalgo et al.	2018	1.24	0.78 – 1.98	17.7
Di Mauro et al.	2017	1.39	1.05 – 1.86	24.4
Total		**1.38**	**1.03 – 1.86**	**100**

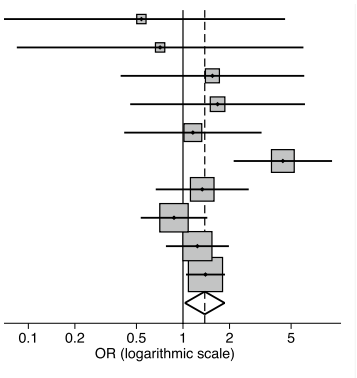
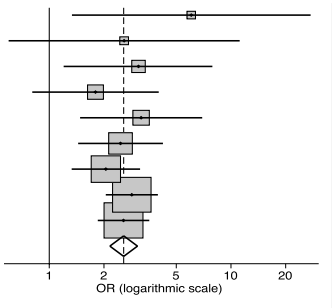
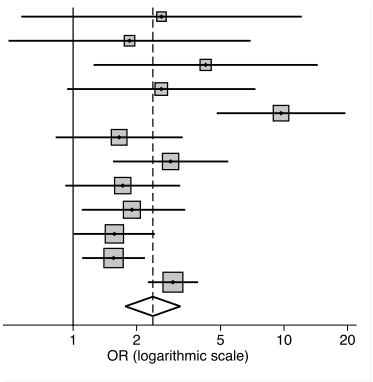


Fig. 3 (continued)

Study	Year	OR	95% CI Lower - Upper	Weights
Oh et al.	2014	6.04	1.33 – 27.43	1.34
Spilopoulos K et al.	2014	2.58	0.59 – 11.17	1.43
Caes et al.	2013	3.1	1.2 – 7.9	3.45
Varela et al.	2017	1.79	0.81 – 4.00	4.77
David et al.	2007	3.2	1.48 – 6.95	5.12
Farag et al.	2017	2.47	1.44 – 4.22	10.6
Olmos et al.	2017	2.05	1.33 – 3.16	16.3
Di Mauro et al.	2017	2.84	2.05 – 3.96	28.1
Fernández-Hidalgo et al.	2018	2.56	1.84 – 3.55	28.8
Total		**2.56**	**2.15 – 3.06**	**100**



Study	Year	OR	95% CI Lower - Upper	Weights
Spilopoulos K et al.	2014	2.63	0.57 – 12.13	3.08
Oh et al.	2014	1.85	0.49 – 6.94	3.87
Hanai et al.	2008	4.25	1.25 – 14.45	4.33
Fayad G. et al.	2011	2.62	0.94 – 7.31	5.49
Perrota et al.	2017	9.68	4.79 – 19.53	8.36
Varela et al.	2017	1.65	0.83 – 3.31	8.45
Hussain et al.	2014	2.89	1.55 – 5.44	9.19
David et al.	2007	1.72	0.92 – 3.22	9.21
Martínez-Sellés et al.	2014	1.9	1.1 – 3.4	9.98
Olmos et al.	2017	1.57	1.01 – 2.44	11.6
Fernández-Hidalgo et al.	2018	1.56	1.10 – 2.19	12.8
Di Mauro et al.	2017	2.97	2.26 – 3.91	13.7
Total		**2.39**	**1.77 – 3.22**	**100**



Study	Year	OR	95% CI Lower - Upper	Weights
Spilopoulos K et al.	2014	2.11	0.46 – 9.66	1.43
Hanai et al.	2008	1.41	0.34 – 5.75	1.66
Oh et al.	2014	2.58	0.71 – 9.42	1.96
Fayad G. et al.	2011	1.22	0.43 – 3.41	3.11
Caes et al.	2013	3.5	1.5 – 8.5	4.37
Varela et al.	2017	1.7	0.77 – 3.78	5.18
Mestres et al.	2007	1.60	0.74 – 3.47	5.53
David et al.	2007	1.85	0.94 – 3.62	7.28
Martínez-Sellés et al.	2014	2.3	1.3 – 4.1	9.98
Olmos et al.	2017	1.49	0.88 – 2.54	11.8
Di Mauro et al.	2017	2.83	2.18 – 3.68	47.7
Total		**2.27**	**1.89 – 2.73**	**100**

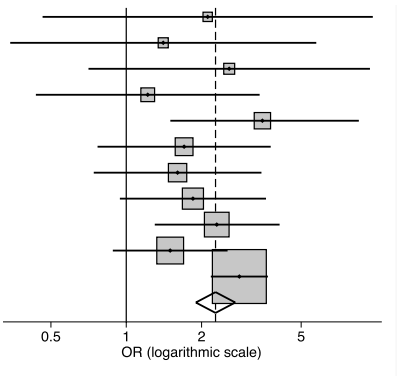


Fig. 3 (continued)

Other data

There was a possible publication bias, according to Egger's method in 7 of the analyzed variables; however, no important asymmetries were observed regarding Funnel plots (Fig. 4).

Regarding the influence analysis, the pooled estimates were robust and independent of the deletion of any individual study.

Discussion

We performed a systematic review of all the studies that reported factors related to in-hospital mortality after cardiac surgery for IE, trying to address which of those previously described factors are related to poor prognosis. After the development of 11 independent meta-analysis, in order to assess the relative influence of each one on mortality, we found that the 11 possible pre-operative prognostic factors analyzed were associated with poor outcome after surgery.

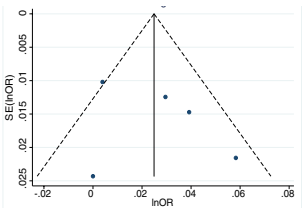
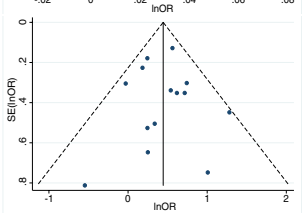
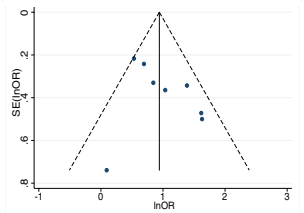
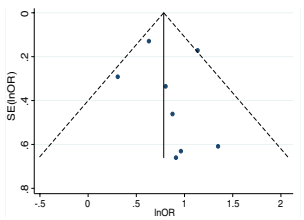
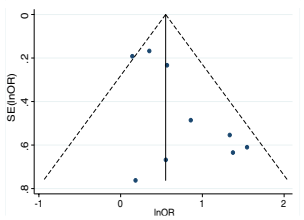
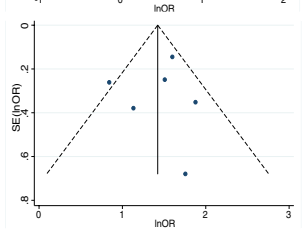
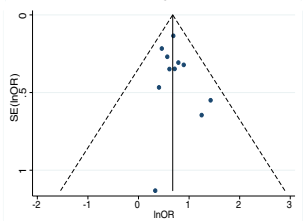
	Funnel Plot	Nº studies	Nº studies ad data	Zero c.	Nº patients	I^2 (%)	Model	Egger test	OR change! (%)	OR (95% CI)	OR (95% CI) Zero c.
1. Age		15	5	-	1393	53.11	REM	0.92	-	1.03 (1.00-1.05)	-
2. Female sex		14	14	0	6326	3.96	FEM	0.02	5.85	1.56 (1.35-1.81)	-
3. Urgent surgery		9	9	1	2809	11.37	REM	0.47	-	2.56 (1.87-3.49)	2.39 (1.91-3.00)
4. Previous cardiac surgery		8	8	0	4796	24.9	FEM	0.01	12.1	2.19 (1.84-2.61)	-
5. NYHA≥ III		10	9	1	5119	32.9	REM	0.83	-	1.73 (1.30-2.31)	1.84 (1.33-2.55)
6. Cardiogenic shock		6	6	0	4627	43.82	REM	0.01	-	4.15 (3.06-5.64)	-
7. Prosthetic valve endocarditis		11	11	0	5857	0	FEM	0.001	4.02	1.98 (1.68-2.33)	-

Fig. 4 Funnel plots and results in each of the meta-analysis conducted. **Nº* number, *Stud ad. data* studies with adequate data for pooled calculations, *FIM* fixed effects model, *REM* random effects model, *OR* odds ratio, *CI* confidence interval, *Zero c* zero correction needed. OR change ^ after influence analysis, maximum change in the pooled OR

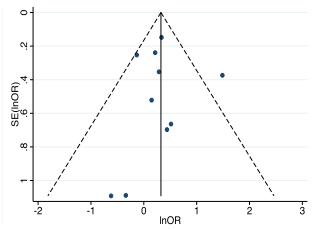
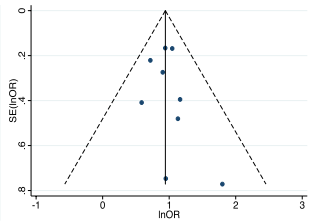
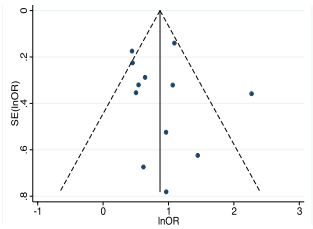
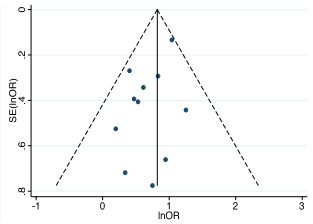
8. Multivalvular involvement		12	10	2	5750	37.92	REM	0.24	-	1.38 (1.03-1.86)	1.35 (1.01-1.82)
9. Renal failure		9	9	0	5267	0	FEM	0.001	4.50	2.57 (2.15-3.06)	-
10. Paravalvular abscess		12	12	0	6422	64.68	REM	0.04	-	2.39 (1.77-3.22)	-
11. *S. aureus*		11	11	0	5759	0	FEM	0.001	18.1	2.27 (1.89-2.73)	-

Fig. 4 (continued)

In the last years, some new endocarditis specific scores have been published [8, 9, 41, 45, 48, 53–55] trying to improve prognostic accuracy. And, at the same time, a lot of retrospective and prospective observational studies with limited sample sizes were published describing independent factors for mortality. Several IE-specific factors have been previously defined as independent mortality predictors in patients with IE [8–10, 47, 56]. However, the impact of these specific factors on postoperative mortality is debated, since the ones considered in each study are different, and with varied effect-sizes.

Regarding age, although it is one of the most important factors employed in risk calculation [41, 44, 47], it was only considered as a continuous variable in 5 of the included studies. There is an important lack of consensus when categorizing this variable, with different cut-off points when age was considered as a categorical variable. Therefore, it is not possible to combine every study´s individual results. Nevertheless, we observed the huge importance that age has in the prognosis after cardiac surgery, with a 3% of increase in mortality for each additional year.

Several studies found that although women suffered from IE less frequently than men, females have more severe manifestations and are more likely to have worse outcomes [57]. Although female sex is considered as a risk factor in the majority of cardiac surgery scores, it is only present in half of the IE-specific scores [47]. The individual studies included in our systematic review did not find female sex as a significant mortality factor, probably because of a relatively small sample size. Therefore, the pooled estimation found that female sex had a significant association with mortality (OR 1.56).

Both IE-specific scores and classic cardiac surgery risk scores take into account the priority of the surgical intervention to estimate mortality risk [47]. Urgent or emergency surgeries had a significant association with mortality in almost every individual study; consequently, the pooled OR was also statistically significant. Similarly, a history of previous cardiac surgery also increases the difficulty of the surgical approach and increases the mortality risk. Probably for that reason, it also showed a significant association with mortality.

Regarding the clinical presentation of IE, a poor preoperative hemodynamic condition is the greatest predictor of mortality, therefore, it is present in all the published risk models [8, 47]. Cardiogenic shock showed the strongest association with mortality in our meta-analysis (OR 4.15). Likewise, in patients without a critical hemodynamic condition, the presence of heart failure also was an indicator of worse prognosis after surgery [5], with a pooled OR for NYHA ≥ III of 1.84. Regarding valvular invasion, on the one hand, the prognostic implications of prosthetic valve IE have been reported elsewhere [58, 59]. On the other hand, the involvement of more than one cardiac valve has been described as a marker of non-controlled infection and greater severity of IE [8]. Although few studies analyzed the risk of death in multivalvular IE, high mortality rates were reported [60, 61]. Our pooled estimations confirmed an association with mortality of both prosthetic valve IE and multivalvular invasion.

Preoperative renal insufficiency is associated with poor prognosis [62, 63] and longer hospitalization in patients with IE [64]. Nine articles considered renal failure as a prognostic factor, and their results were consistent. Paravalvular abscess increases complexity of the procedure and complications associated with IE surgery and mortality [13]. It is probably associated with higher microorganism virulence, poor prognosis and a delay in the surgical treatment [65]. Twelve studies analyzed paravalvular abscesses, and the pooled OR was 2.39, which agrees with the studies that specifically analyzed outcomes in IE complicated with paravalvular abscesses [13, 65]. *Staphylococcus aureus* is being recognized as the most common cause of IE [1, 66] and also has been shown to be an independent predictor of mortality. *Staphylococcus aureus* is related to severity of the local invasion, with increased likelihood of abscesses formation, fistulae and prosthesis dehiscence. Our results confirmed a significant relation with mortality of this etiologic agent.

We believe that an adequate knowledge of the impact on survival of the analyzed risk factors may help to guide the surgical decision process in IE patients. Some meta-analyses were developed to assess the optimal time to surgery, compared biological versus mechanical prosthesis, or compared valve replacement versus valve repair in IE patients. But to our knowledge, our study is the first meta-analysis of prognostic factors in IE surgery in which 11 factors associated with poor outcome were assessed. Recently, Wang et al. published a meta-analysis of risk-scores in surgery for IE [67], comparing the prognostic utility of EuroSCORE I and II, which included 8 studies (1743 patients), and they calculated pooled c-statistics for operative mortality for both scores. They concluded that EuroSCORE had a trend to over-estimate mortality, and suggested that EuroSCORE II would be a better estimator, but in addition, the authors made great emphasis on the need to develop new endocarditis-specific risk scores. One possible future study, after identification of these preoperative factors related to in-hospital mortality, could be the use of the obtained regression coefficients in the development of a new specific-IE score derived from the literature search.

Study limitations

The major limitation of this report is that our meta-analysis included both prospective and retrospective observational studies, with a small to moderate number of patients. Therefore, because of their observational nature, there could be some unidentified differences between the studies, regarding different disease spectrum, referring population size, epidemiological factors, hospital differences, referral bias [68] or surgical indications. For example, the aggressiveness of endocarditis could be different between the included studies, since some of them included only left site IE; whereas other studies included also right side IE. In addition, some studies considered in-hospital mortality as their main outcome, whereas other calculated 30-day mortality. However, predefined criteria insured the inclusion of exclusively active IE. Unfortunately, the results of the individual studies included were not adjusted for baseline differences, since we reported the OR from the univariate analysis or the ones obtained from raw data calculations; so there is risk of bias because of unknown confounders. Of the 16 studies finally included in the meta-analysis, it is important to note that none of them included as much as these 11 factors altogether.

Statistical analysis revealed a possible publication bias in some of the analyzed factors. In addition, some of the funnel plot showed slight asymmetries. The role of chance is critical for interpretation of funnel plots because most of our meta-analyses contained few studies. Relations across studies in meta-analysis are seriously prone to false positive findings when there is heterogeneity and a few numbers of studies are included, which may affect funnel plot symmetry [33].

In reference to age, as we previously mentioned, there is an important lack of consensus in the cut-off points for categorization of age across the studies. Categorization of continuous variables carries a loss of information, and it is not recommended. Therefore, it could only be analyzed when continuous age OR were reported.

All those variables that were considered as risk factors for mortality in two or more IE-specific scores were included in the meta-analysis; however, some other possible risk factors were not analyzed, such as vegetations, thrombocytopenia, LVEF, stroke or embolism. We tried to follow a strict criterion to choose the most studied variables in risk assessment by reviewing previously published IE-specific scores; however the analysis of those other risk factors could be performed to further improve mortality risk prediction.

Conclusions

After a systematic review, we identified 11 preoperative factors related with an increased postoperative mortality: Cardiogenic shock, urgent surgery, paravalvular abscess, preoperative renal failure, previous cardiac surgery, *S. aureus*, female sex, age, NHYA class ≥ III, prosthetic valve IE, and multivalvular involvement.

The meta-analysis of each of these factors showed a significant association with an increased in-hospital mortality after surgery for active infective endocarditis.

Acknowledgements The present authors sincerely thank Noelia Álvarez for his contribution to the literature search.

Funding This research received no specific grant from any funding agency, commercial or non-profit.

Compliance with ethical standards

Conflict of interest There was no conflict of interest.

Ethical approval The Ethical Review Board (ERB) of the Hospital approved the implementation of this study (ERB number 313/2016, approved on November 28, 2016).

Informed consent The requirement for informed written consent was waived. Patient identification was encoded, complying with the requirements of the Organic Law on Data Protection 15/1999.

References

1. Murdoch DR, Corey GR, Hoen B, Miró JM, Fowler VG, Bayer AS, et al. Clinical presentation, etiology, and outcome of infective endocarditis in the 21st century: the International Collaboration on Endocarditis-Prospective Cohort Study. Arch Intern Med. 2009;169:463–73.
2. San Román JA, López J, Vilacosta I, Luaces M, Sarriá C, Revilla A, et al. Prognostic stratification of patients with left-sided endocarditis determined at admission. Am J Med. 2007;120:369.e1–7.
3. Rasmussen RV, Bruun LE, Lund J, Larsen CT, Hassager C, Bruun NE. The impact of cardiac surgery in native valve infective endocarditis: can euroSCORE guide patient selection? Int J Cardiol. 2011;149:304–9.
4. Chu VH, Park LP, Athan E, Delahaye F, Freiberger T, Lamas C, et al. Association between surgical indications, operative risk, and clinical outcome in infective endocarditis a prospective study from the international collaboration on endocarditis. Circulation. 2015;131:131–40.
5. Prendergast BD, Tornos P. Surgery for infective endocarditis: who and when? Circulation. 2010;121:1141–52.
6. Gatti G, Chocron S, Obadia J-F, Duval X, Iung B, Alla F, et al. Using surgical risk scores in nonsurgically treated infective endocarditis patients. Hell J Cardiol HJC Hell Kardiologike Epitheorese. 2019. https://doi.org/10.1016/j.hjc.2019.01.008 **(Epub ahead of print)**.
7. Iung B, Doco-Lecompte T, Chocron S, Strady C, Delahaye F, Le Moing V, et al. Cardiac surgery during the acute phase of infective endocarditis: discrepancies between European Society of Cardiology guidelines and practices. Eur Heart J. 2016;37:840–8.
8. Gaca JG, Sheng S, Daneshmand MA, O'Brien S, Rankin JS, Brennan JM, et al. Outcomes for endocarditis surgery in North America: a simplified risk scoring system. J Thorac Cardiovasc Surg. 2011;141:98–106.
9. Olmos C, Vilacosta I, Habib G, Maroto L, Fernández C, López J, et al. Risk score for cardiac surgery in active left-sided infective endocarditis. Heart. 2017;103:1435–42.
10. Farag M, Borst T, Sabashnikov A, Zeriouh M, Schmack B, Arif R, et al. Surgery for infective endocarditis: outcomes and predictors of mortality in 360 consecutive patients. Med Sci Monit Int Med J Exp Clin Res. 2017;23:3617–26.
11. Ramos-Martínez A, Calderón-Parra J, Miró JM, Muñoz P, Rodríguez-Abella H, Valerio M, et al. Effect of the type of surgical indication on mortality in patients with infective endocarditis who are rejected for surgical intervention. Int J Cardiol. 2019. https://doi.org/10.1016/j.ijcard.2019.01.014 **(Epub ahead of print)**.
12. Varela Barca L, López-Menéndez J, Navas Elorza E, Moya Mur JL, Centella Hernéndez T, Redondo Palacios A, et al. Long-term prognosis after surgery for infective endocarditis: distinction between predictors of early and late survival. Enfermedades Infecc Microbiol Clín. 2018. https://doi.org/10.1016/j.eimc.2018.10.017 **(Epub ahead of print)**.
13. Varela Barca L, López Menéndez J, Martín García M, Redondo Palacios A, Centella Hernández T, Miguelena Hycka J, et al. Absceso paravalvular en la endocarditis bacteriana: influencia en el pronóstico postoperatorio. Circ Cardiovasc. 2017;24:2–7.
14. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLOS Med. 2009;6:e1000097.
15. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA. 2000;283:2008–12.
16. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356:i6460.
17. Amilivia G. Infective endocarditis: a better outcome after surgery during the active phase. Crit Care. 2000;2:6339.
18. Horstkotte D, Follath F, Gutschik E, Lengyel M, Oto A, Pavie A, et al. Guidelines on prevention, diagnosis and treatment of infective endocarditis executive summary; the task force on infective endocarditis of the European society of cardiology. Eur Heart J. 2004;25:267–76.
19. Pettersson GB. Surgical treatment of endocarditis. Tex Heart Inst J. 2011;38:667–8.
20. Sexton DJ, Spelman D. Current best practices and guidelines. Assessment and management of complications in infective endocarditis. Cardiol Clin. 2003;21:273–82.
21. Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med. 2013;158:280–6.
22. Clark RE. Definitions of terms of the society of thoracic surgeons national cardiac surgery database. Ann Thorac Surg. 1994;58:271–3.
23. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur Heart J. 2016;37:2129–200.
24. Hollenberg SM, Kavinsky CJ, Parrillo JE. Cardiogenic shock. Ann Intern Med. 1999;131:47–59.

25. Thomas D, Desruennes M, Jault F, Isnard R, Gandjbakhch I. Cardiac and extracardiac abscesses in bacterial endocarditis. Arch Mal Coeur Vaiss. 1993;86:1825–35.
26. Edmunds LH, Clark RE, Cohn LH, Grunkemeier GL, Miller DC, Weisel RD. Guidelines for reporting morbidity and mortality after cardiac valvular operations. The American Association for Thoracic Surgery, Ad Hoc Liaison Committee for Standardizing Definitions of Prosthetic Heart Valve Morbidity. Ann Thorac Surg. 1996;62:932–5.
27. Akins CW, Miller DC, Turina MI, Kouchoukos NT, Blackstone EH, Grunkemeier GL, et al. Guidelines for reporting mortality and morbidity after cardiac valve interventions. Ann Thorac Surg. 2008;85:1490–5.
28. Doménech JM (2018) Meta-analysis OR, RR, RD, IR, ID, B, MD & R combined: user-written command mar for Stata [computer program]. V1.4.1. Barcelona: Graunt 21.
29. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327:557–60.
30. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ. 2011;342:d549.
31. Barili F, Parolari A, Kappetein PA, Freemantle N. Statistical primer: heterogeneity, random- or fixed-effects model analyses? Interact Cardiovasc Thorac Surg. 2018;27:317–21.
32. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315:629–34.
33. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011;343:d4002.
34. Rostagno C, Rosso G, Puggelli F, Gelsomino S, Braconi L, Montesi GF, et al. Active infective endocarditis: clinical characteristics and factors related to hospital mortality. Cardiol J. 2010;17:566–73.
35. Mestres CA, Castro MA, Bernabeu E, Josa M, Cartaná R, Pomar JL, et al. Preoperative risk stratification in infective endocarditis. Does the EuroSCORE model work?Preliminary results. Eur J Cardio-Thorac Surg. 2007;32:281–5.
36. Hanai M, Hashimoto K, Mashiko K, Sasaki T, Sakamoto Y, Shiratori K, et al. Active infective endocarditis: management and risk analysis of hospital death from 24 years' experience. Circ J. 2008;72:2062–8.
37. Fayad G, Leroy G, Devos P, Hervieux E, Senneville E, Koussa M, et al. Characteristics and prognosis of patients requiring valve surgery during active infective endocarditis. J Heart Valve Dis. 2011;20:223–8.
38. Meszaros K, Nujic S, Sodeck GH, Englberger L, König T, Schönhoff F, et al. Long-term results after operations for active infective endocarditis in native and prosthetic valves. Ann Thorac Surg. 2012;94:1204–10.
39. Caes F, Bove T, Van Belleghem Y, Vandenplas G, Van Nooten G, Francois K. Reappraisal of a single-centre policy on the contemporary surgical management of active infective endocarditis. Interact Cardiovasc Thorac Surg. 2014;18:169–76.
40. Hussain ST, Shrestha NK, Gordon SM, Houghtaling PL, Blackstone EH, Pettersson GB. Residual patient, anatomic, and surgical obstacles in treating active left-sided infective endocarditis. J Thorac Cardiovasc Surg. 2014;148:981–8.
41. Martínez-Sellés M, Muñoz P, Arnáiz A, Moreno M, Gálvez J, Rodríguez-Roda J, et al. Valve surgery in active infective endocarditis: a simple score to predict in-hospital prognosis. Int J Cardiol. 2014;175:133–7.
42. Spiliopoulos K, Giamouzis G, Haschemi A, Karangelis D, Antonopoulos N, Fink G, et al. Surgical management of infective endocarditis: early and long-term mortality analysis. Single-center experience and brief literature review. Hellenic J Cardiol. 2014;55:462–74.
43. Marks DJB, Hyams C, Koo CY, Pavlou M, Robbins J, Koo CS, et al. Clinical features, microbiology and surgical outcomes of infective endocarditis: a 13-year study from a UK tertiary cardiothoracic referral centre. QJM Mon J Assoc Physicians. 2015;108:219–29.
44. Oh T, Voss J, Gamble G, Kang N, Pemberton J. Comparison of contemporary risk scores for predicting outcomes after surgery for active infective endocarditis. Heart Vessels. 2015;30:227–34.
45. Di Mauro M, Dato GMA, Barili F, Gelsomino S, Sante P, Corte AD, et al. A predictive model for early mortality after surgical treatment of heart valve or prosthesis infective endocarditis. The EndoSCORE. Int J Cardiol. 2017;241:97–102.
46. Perrotta S, Jeppsson A, Frojd V, Svensson G. Surgical treatment for infective endocarditis: a single-centre experience. Thorac Cardiovasc Surg. 2017;65:166–73.
47. Varela L, López-Menéndez J, Redondo A, Fajardo ER, Miguelena J, Centella T, et al. Mortality risk prediction in infective endocarditis surgery: reliability analysis of specific scores. Eur J Cardio-Thorac Surg. 2018;53:1049–54.
48. Fernández-Hidalgo N, Ferreria-González I, Marsal JR, Ribera A, Aznar ML, de Alarcón A, et al. A pragmatic approach for mortality prediction after surgery in infective endocarditis: optimizing and refining EuroSCORE. Clin Microbiol Infect. 2018;24:1102.
49. David TE, Gavra G, Feindel CM, Regesta T, Armstrong S, Maganti MD. Surgical treatment of active infective endocarditis: a continued challenge. J Thorac Cardiovasc Surg. 2007;133:144–9.
50. Ferrera C, Vilacosta I, Fernández C, López J, Sarriá C, Olmos C, et al. Early surgery for acute-onset infective endocarditis. Eur J Cardio-Thorac Surg. 2018;54:1060–6.
51. Garcia-Granja PE, Lopez J, Vilacosta I, Ortiz-Bautista C, Sevilla T, Olmos C, et al. Polymicrobial infective endocarditis: clinical features and prognosis. Medicine. 2015;94:e2000. https://doi.org/10.1097/md.0000000000002000.
52. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928.
53. De Feo M, Cotrufo M, Carozza A, De Santo LS, Amendolara F, Giordano S, et al. The need for a specific risk prediction system in native valve infective endocarditis surgery. Sci World J. 2012;2012:307571. https://doi.org/10.1100/2012/307571.
54. Gatti G, Perrotti A, Obadia J-F, Duval X, Iung B, Alla F, et al. Simple scoring system to predict in-hospital mortality after surgery for infective endocarditis. J Am Heart Assoc. 2017;6:e004806.
55. da Costa MAC, Wollmann DR, Campos ACL, da Cunha CLP, de Carvalho RG, de Andrade DF, et al. Risk index for death by infective endocarditis: a multivariate logistic model. Rev Bras Cir Cardiovasc. 2007;22:192–200.
56. Nunes MCP, Guimarães-Júnior MH, Murta Pinto PHO, Coelho RMP, Souza Barros TL, de Faleiro Maia NPA, et al. Outcomes of infective endocarditis in the current era: early predictors of a poor prognosis. Int J Infect Dis. 2018;68:102–7.
57. Sambola A, Fernández-Hidalgo N, Almirante B, Roca I, González-Alujas T, Serra B, et al. Sex differences in native-valve infective endocarditis in a single tertiary-care hospital. Am J Cardiol. 2010;106:92–8.
58. Nagy M, Alkady H, Abo Senna W, Abdelhay S. Predictors of surgical outcome in isolated prosthetic mitral valve endocarditis. Asian Cardiovasc Thorac Ann. 2018;26:517–23.
59. Lopez J, Sevilla T, Vilacosta I, Garcia H, Sarria C, Pozo E, et al. Clinical significance of congestive heart failure in prosthetic

valve endocarditis. A multicenter study with 257 patients. Rev Espanola Cardiol. 2013;66:384–90.
60. Mueller XM, Tevaearai HT, Stumpe F, Fischer AP, Hurni M, Ruchat P, et al. Multivalvular surgery for infective endocarditis. Cardiovasc Surg. 1999;7:402–8.
61. Kim N, Lazar JM, Cunha BA, Liao W, Minnaganti V. Multivalvular endocarditis. Clin Microbiol Infect. 2000;6:207–12.
62. Chen C-H, Lo M-C, Hwang K-L, Liu C-E, Young T-G. Infective endocarditis with neurologic complications: 10-year experience. J Microbiol Immunol Infect. 2001;34:119–24.
63. Liu Y, Zhang H, Liu Y, Han Q, Tang Y, Zhao L, et al. Risk factors and short-term prognosis of preoperative renal insufficiency in infective endocarditis. J Thorac Dis. 2018;10:3679–88.
64. Tamura K, Arai H, Yoshizaki T. Long-term outcome of active infective endocarditis with renal insufficiency in cardiac surgery. Ann Thorac Cardiovasc Surg. 2012;18:216–21.
65. David TE, Regesta T, Gavra G, Armstrong S, Maganti MD. Surgical treatment of paravalvular abscess: long-term results. Eur J Cardio-Thorac Surg. 2007;31:43–8.
66. Fernández-Hidalgo N, Ribera A, Larrosa MN, Viedma E, Origüen J, de Alarcón A, et al. Impact of Staphylococcus aureus phenotype and genotype on the clinical characteristics and outcome of infective endocarditis. A multicentre, longitudinal, prospective, observational study. Clin Microbiol Infect. 2018;24:985–91.
67. Wang TKM, Wang MTM, Pemberton J. Risk scores and surgery for infective endocarditis: a meta-analysis. Int J Cardiol. 2016;222:1001–2.
68. Kanafani ZA, Kanj SS, Cabell CH, Cecchi E, de Oliveira Ramos A, Lejko-Zupanc T, et al. Revisiting the effect of referral bias on the clinical spectrum of infective endocarditis in adults. Eur J Clin Microbiol Infect Dis. 2010;29:1203–10.

European Journal of Cardio-Thoracic Surgery 0 (2019) 1–8
doi:10.1093/ejcts/ezz328

ORIGINAL ARTICLE

Cite this article as: Varela Barca L, Fernández-Felix BM, Navas Elorza E, Mestres CA, Muñoz P, Cuerpo-Caballero G *et al.* Prognostic assessment of valvular surgery in active infective endocarditis: multicentric nationwide validation of a new score developed from a meta-analysis. Eur J Cardiothorac Surg 2019; doi:10.1093/ejcts/ezz328.

Prognostic assessment of valvular surgery in active infective endocarditis: multicentric nationwide validation of a new score developed from a meta-analysis

Laura Varela Barca [a,b,*], Borja M. Fernández-Felix[b,c], Enrique Navas Elorza [d], Carlos A. Mestres [e], Patricia Muñoz [f], Gregorio Cuerpo-Caballero [g], Hugo Rodríguez-Abella[g], Miguel Montejo-Baranda[h], Regino Rodríguez-Álvarez[i], Francisco Gutiérrez Díez[j], Miguel Angel Goenaga[j], Eduard Quintana[k], Guillermo Ojeda-Burgos [l], Arístides de Alarcón[m], Laura Vidal-Bonet[a], Tomasa Centella Hernández[b,n] and Jose López-Menéndez [b,n], on behalf of the Spanish Collaboration on Endocarditis–Grupo de Apoyo al Manejo de la Endocarditis infecciosa en ESpaña (GAMES)[†]

[a] Department of Cardiovascular Surgery, University Hospital Son Espases, Palma de Mallorca, Spain
[b] University of Alcalá de Henares, Madrid, Spain
[c] CIBER Epidemiology and Public Health (CIBERESP), Clinical Biostatistics Unit, Hospital Ramon y Cajal (IRYCIS), Madrid, Spain
[d] Department of Infectology, University Hospital Ramon y Cajal, Madrid, Spain
[e] Department of Cardiovascular Surgery, University Hospital Zurich, Zurich, Switzerland
[f] Department of Clinical Microbiology and Infectious Diseases, University Hospital Gregorio Marañón, Madrid, Spain
[g] Department of Cardiovascular Surgery, University Hospital Gregorio Marañón, Madrid, Spain
[h] Department of Infectology, University Hospital Cruces, Bilbao, Spain
[i] Department of Cardiovascular Surgery, University Hospital Marques de Valdecilla, Santander, Spain
[j] Department of Infectology, University Hospital Donosti, San Sebastian, Spain
[k] Department of Cardiovascular Surgery, Hospital Clínic de Barcelona, University of Barcelona, Barcelona, Spain
[l] Department of Infectology, University Hospital Virgen de la Victoria, Malaga, Spain
[m] Clinical Unit of Infectious Diseases, Microbiology, and Preventive Medicine, Infectious Diseases Research Group, Institute of Biomedicine of Seville (IBiS), University of Seville, CSIC, University Hospital Virgen del Rocío, Seville, Spain
[n] Department of Cardiovascular Surgery, University Hospital Ramon y Cajal, Madrid, Spain

* Corresponding author. Department of Cardiovascular Surgery, University Hospital Son Espases, Ctra de Valldemossa, 79, 07120 Palma, Illes Balears, Spain. Tel: +34617103918; e-mail: lauravarela21089@gmail.com (L.V. Barca).

Received 17 July 2019; received in revised form 24 October 2019; accepted 27 October 2019

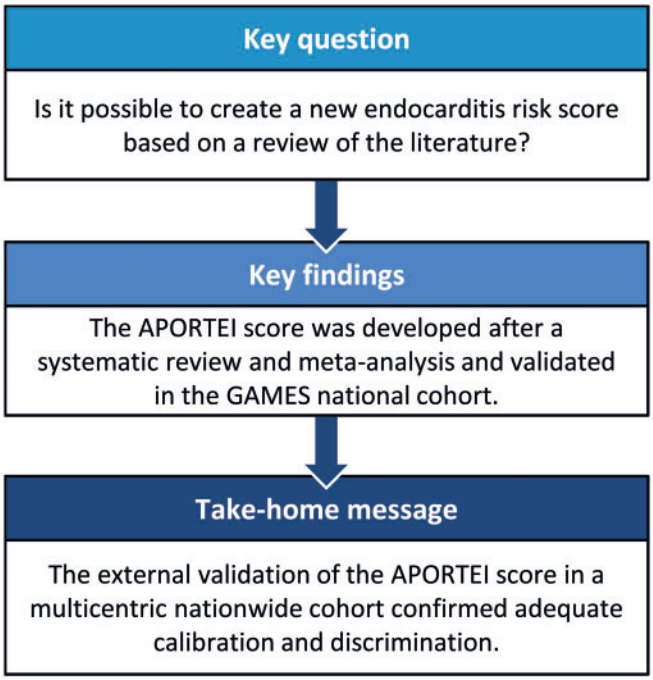
Abstract

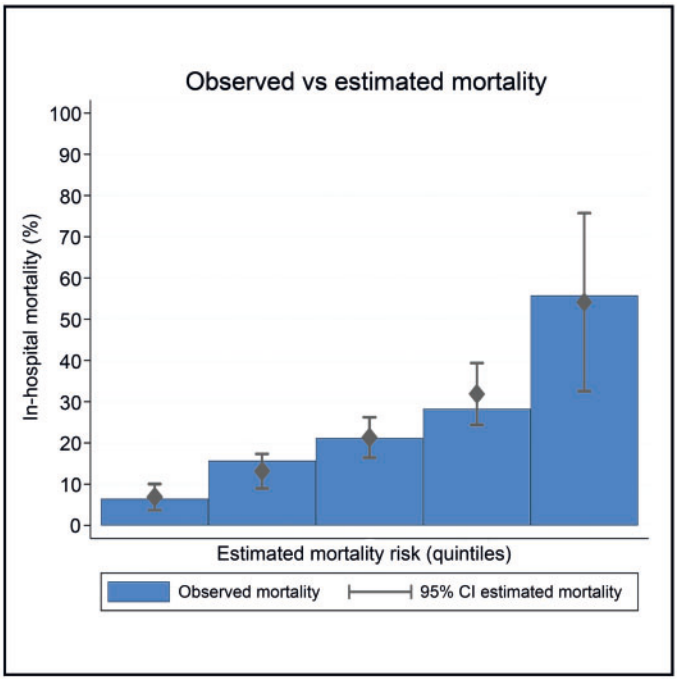
OBJECTIVES: Several risk prediction models have been developed to estimate the risk of mortality after valve surgery for active infective endocarditis (IE), but few external validations have been conducted to assess their accuracy. We previously developed a systematic review and meta-analysis of the impact of IE-specific factors for the in-hospital mortality rate after IE valve surgery, whose obtained pooled estimations were the basis for the development of a new score (APORTEI). The aim of the present study was to assess its prognostic accuracy in a nationwide cohort.

METHODS: We analysed the prognostic utility of the APORTEI score using patient-level data from a multicentric national cohort. Patients who underwent surgery for active IE between 2008 and 2018 were included. Discrimination was evaluated using the area under the receiver operating characteristic curve, and the calibration was assessed using the calibration slope and the Hosmer–Lemeshow test. Agreement between the APORTEI and the EuroSCORE I was also analysed by Lin's concordance correlation coefficient (CCC), the Bland–Altman agreement analysis and a scatterplot graph.

RESULTS: The 11 variables that comprised the APORTEI score were analysed in the sample. The APORTEI score was calculated in 1338 patients. The overall observed surgical mortality rate was 25.56%. The score demonstrated adequate discrimination (area under the receiver operating characteristic curve = 0.75; 95% confidence interval 0.72–0.77) and calibration (calibration slope = 1.03; Hosmer–Lemeshow test $P = 0.389$). We found a lack of agreement between the APORTEI and EuroSCORE I (concordance correlation coefficient = 0.55).

Presented at the 33rd Annual Meeting of the European Association for Cardio-Thoracic Surgery, Lisbon, Portugal, 3–5 October 2019.
†Members are listed in the Supplementary Material S1.





CONCLUSIONS: The APORTEI score, developed from a systematic review and meta-analysis, showed an adequate estimation of the risk of mortality after IE valve surgery in a nationwide cohort.

Keywords: Infective endocarditis • Valve surgery • Risk score • Systematic review and meta-analysis

ABBREVIATIONS

AUC	Area under the receiver operating characteristic curve
CI	Confidence interval
HLt	Hosmer–Lemeshow test
IESF	IE-specific factors
IE	Infective endocarditis
STS	Society of Thoracic Surgeons
SD	Standard deviation

INTRODUCTION

A large proportion of patients with active infective endocarditis (IE) are at high risk of death without cardiovascular surgery [1]. Currently available classical risk scores have a poor performance in predicting mortality in those patients with an established indication for heart valve surgery.

Several IE-specific factors (IESF) were defined as independent predictors of mortality in patients with IE [2]. Since then, some new IE-specific scores have been published [2–9] with the aim of improving the prognostic accuracy in the surgical treatment of IE. These specific scores incorporate those IESF that are considered to be associated with postoperative mortality, and they have demonstrated a more accurate prediction of mortality in comparison with the classical scores [3, 10]. Nevertheless, there is a lack of consensus as to which IESF really impact the mortality rate and as to the weight of the association of each factor with a poor outcome. The utility of these risk models in clinical practice has been debated [11, 12].

A few meta-analyses and systematic reviews about the prognosis of IE have been developed, but they are only applicable to specific subgroups of patients [12, 13]. We recently conducted a systematic review and meta-analysis of the impact of these IESF on the prognosis after cardiac surgery [14] by following the PRISMA guidelines [15], which represent the results of the combination of 16 studies, including 7484 patients. As a result, we identified 11 IESF with significant impact on the mortality rate. The regression coefficients of the pooled estimates were used to construct a new scoring system.

The newly created score has been named the APORTEI score, which stands for '*A*nálisis de los factores *PRO*nósticos en el *T*ratamiento quirúrgico de la *E*ndocarditis *I*nfecciosa'. An online calculator was created to facilitate the use of APORTEI (Supplementary Material S2).

The aim of the present study was to development and validate this new score in a national cohort of patients with the diagnosis of IE.

METHODS

Study design

The prognostic accuracy of this newly created score (APORTEI score) was validated in a nationwide sample (GAMES cohort). That sample included all the IE episodes registered in the Spanish endocarditis cohort affecting adult subjects (older than 18 years)

Table 1: Score development

Predictor	Studies[a]	OR (95% CI)	Regression coefficient	Scoring points
Age (years)	15	1.03 (1.00–1.05)	0.03	0.5 × (age-50)[b]
Female gender	14	1.56 (1.35–1.81)	0.44	7
Urgent surgery	9	2.39 (1.91–3.00)	0.87	15
Previous cardiac surgery	8	2.19 (1.84–2.61)	0.78	13
NYHA functional class ≥III	10	1.84 (1.33–2.55)	0.61	10
Cardiogenic shock	6	4.15 (3.06–5.64)	1.42	24
Prosthetic valve	11	1.98 (1.68–2.33)	0.68	11
Multivalvular	12	1.35 (1.01–1.82)	0.30	5
Renal failure	9	2.57 (2.15–3.06)	0.94	16
Abscess	12	2.39 (1.77–3.22)	0.87	15
Staphylococcus aureus	11	2.27 (1.89–2.73)	0.82	14

[a]Number of studies in the meta-analysis that included each factor.
[b]Only if patient age is ≥50 years.
CI: confidence interval; NYHA: New York Heart Association; OR: odds ratio.

who were operated on between 2008 and 2018. All the included patients had a preoperative diagnosis of active IE, and they were treated with cardiac surgery after a multidisciplinary consensus.

Definitions

- GAMES stands for 'Grupo de Apoyo al Manejo de la Endocarditis infecciosa en España' or 'Spanish Collaboration on Endocarditis', which includes patients from 32 different hospitals across Spain. GAMES was activated 1 January 2008 and prospectively enrols patients in a nationwide registry. Regional and local ethics committees approved the study, and patients gave their informed consent.
- The diagnosis of IE was based on Dukés criteria [16], and active IE was defined as an ongoing active infection under antimicrobial treatment at the time of surgery [17]. Cases of IE affecting native and prosthetic valves were included. IE exclusively related to cardiac implantable electronic devices or IE episodes in patients who suffered from previous congenital heart disease and patients younger than 18 years were excluded.
- The end point event used to assess the performance of the model was the postoperative mortality rate, considered as in-hospital death or death during the first 30 days after surgery, according to the surgical guidelines [18, 19].

Definition of the IESF analysed in the sample [14]:

- Urgent surgery: surgery required within 24 h of its indication [20].
- Emergency surgery: surgery required on the day of admission [20].
- Previous cardiac surgery: previous surgical procedure with opening of the pericardium.
- New York Heart Association functional class ≥III.
- Cardiogenic shock: acute myocardial dysfunction, with systolic pressure <90 mmHg, tissue hypoperfusion and low cardiac output [20].
- Prosthetic valve IE: IE affecting a previously inserted prosthetic valve.
- Multivalvular involvement: IE affecting more than 1 heart valve.
- Renal failure: presence of a serum creatinine concentration >2 mg/dl.
- Paravalvular abscess: purulent cavity with necrosis and capacity to invade adjacent structures [19].

APORTEI score development

The weight of each variable included in the model was obtained by dividing the regression coefficient of each variable by the smallest regression coefficient, rounded to the nearest integer [14]. Therefore, the regression coefficients were transformed into risk points to create the new score (Table 1).

Age was considered a continuous variable, and 50 years was considered the lowest cut-point, with increasing score attributed to each additional year over this limit.

Statistical analysis

Continuous variables were expressed as mean and standard deviation (SD) if normally distributed or as median and interquartile range in the presence of marked asymmetries. Normal distribution was assessed using the Shapiro–Wilk test and standardized normal probability plots. Categorical variables were expressed as absolute and relative frequencies.

We assessed the performance of the model through the calculation of its discrimination and calibration over the postoperative mortality rate. Discrimination, the capacity of the score to predict death, was assessed by calculating the area under the receiver operating characteristic curve (AUC). Models are considered to have an optimal performance when the AUC exceeds 0.7. Calibration, defined as the agreement between predicted and observed probabilities, was assessed using the calibration slope. A well-calibrated model has a calibration slope of 1. In addition, we used the Hosmer–Lemeshow test (HLt) goodness-of-fit test after dividing the sample into deciles of risk.

The APORTEI score was used to calculate the estimated probability of hospital mortality. We carried out a logistic regression, with the event 'postoperative mortality' as the dependent variable and the scoring points as the independent term. It was expressed

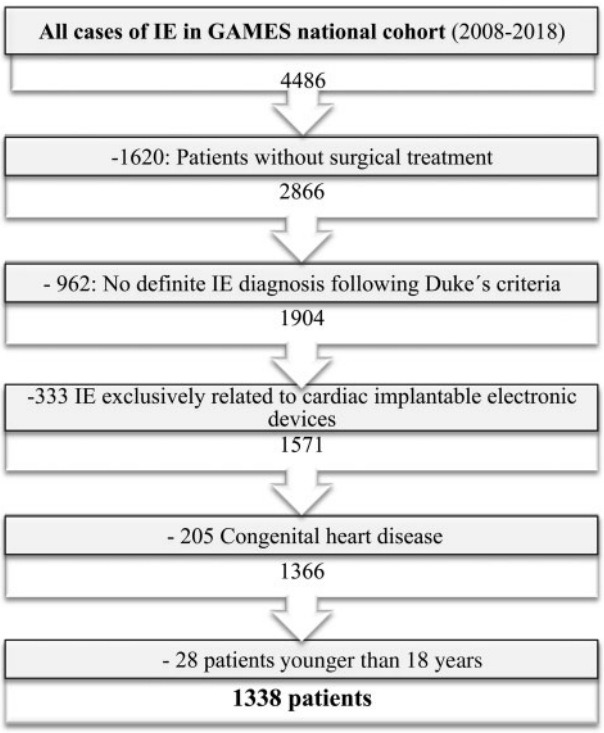


Figure 1: Flow chart for patient selection. IE: infective endocarditis.

as percentages. The Lin's concordance correlation coefficient was calculated to assess the agreement between the estimated mortality rates of using the APORTEI score and the EuroSCORE I, which ranged from 0 to 1, with good agreement at values over 0.9. In addition, the Bland–Altman plot was presented to graphically depict agreement between both scores.

Statistical analyses were performed using Stata/IC 14.2 (Stata Statistical Software: Release 14; StataCorp LP, College Station, TX, USA).

RESULTS

Baseline characteristics

The GAMES national cohort comprised 4486 patients including patients with and without indications for surgery. Of these, 1338 patients underwent valve surgery for active IE between 2008 and 2018 (Fig. 1). The mean age of the cohort of patients who underwent cardiac surgery was 63.6 years (SD 13.2 years); 975 (72.9%) were men. The involved valve was a native valve in 65.6%, and a prosthetic valve in the remaining 34.4%. The observed in-hospital mortality rate was 25.6%. The median logistic EuroSCORE I was 17.7% (interquartile range 6.8–38.2%). The distribution of the 11 IESF in this cohort is shown in Table 2.

APORTEI score assessment

The APORTEI score was calculated for every patient in the GAMES cohort, using individual patient-level data. The AUC was 0.75 [95% confidence interval (CI) 0.72–0.77], the calibration slope was 1.03 and the HLt *P*-value was 0.389. Calibration was visually evaluated using a calibration plot (Fig. 2).

Table 2: Distribution of factors in our sample and comparison with the meta-analysis combining the results

Predictor	Total (*n* = 1338)	OR cohort	OR meta-analysis
Age (years), mean (SD)	63.6 (13.1)	1.04	1.03
Female gender, *n* (%)	363 (27.1)	1.39	1.56
Urgent surgery, *n* (%)	483 (36.1)	2.09	2.39
Previous cardiac surgery, *n* (%)	483 (36.1)	1.95	2.19
NYHA functional class >III, *n* (%)	556 (41.5)	2.15	1.84
Cardiogenic shock, *n* (%)	401 (30.0)	3.44	4.15
Prosthetic valve, *n* (%)	460 (34.4)	2.14	1.95
Multivalvular involvement, *n* (%)	192 (14.3)	1.13	1.35
Renal failure (Cr > 2 mg/dl), *n* (%)	566, (42.3)	2.42	2.57
Abscess, *n* (%)	379 (28.3)	1.48	2.39
Staphylococcus aureus, *n* (%)	231 (17.3)	2.38	2.27

Cr: creatinine; NYHA: New York Heart Association; OR: odds ratio; SD: standard deviation.

The predicted risk of mortality associated with individual scores was calculated, according to the presence and scoring of each IESF. The minimum total score points was 0 and the maximum registered score was 142. The predicted probability of postoperative mortality ranged from 4.8% to 88%.

Prognostic groups were defined based on the observed mortality rate, according to the APORTEI scoring points (Fig. 3).

Agreement between the APORTEI score and the logistic EuroSCORE I

We analysed the prognostic accuracy of the EuroSCORE I in our sample. The AUC was 0.72 (95% CI 0.69–0.75); however, the calibration was suboptimal (HLt $P = 0.062$).

The EuroSCORE I underestimated the mortality rate in low-risk patients and overestimated the mortality rate in high-risk patients. Agreement between the APORTEI and the EuroSCORE I was assessed using the Bland–Altman method (Fig. 4). The difference between the mortality rates predicted by both scores was close to 0 (95% CI -0.02 to 0.01). However, absolute agreement between scores was low (concordance correlation coefficient = 0.55); there was a clear trend towards lack of agreement between both scores in the Bland–Altman graph analysis: the higher the expected mortality rate, the lower was the agreement.

Figure 5 shows the predicted mortality rates of the APORTEI score and the EuroSCORE I, showing graphically the low agreement between both scores with the observed mortality rates in each group.

DISCUSSION

Regarding our results, the APORTEI score demonstrated good discrimination and calibration after the assessment of its prognostic accuracy in the GAMES national cohort.

Infective endocarditis-specific factors

The high mortality rate associated with surgery, reported to range from 15% to more than 45% [1], could be attributed to

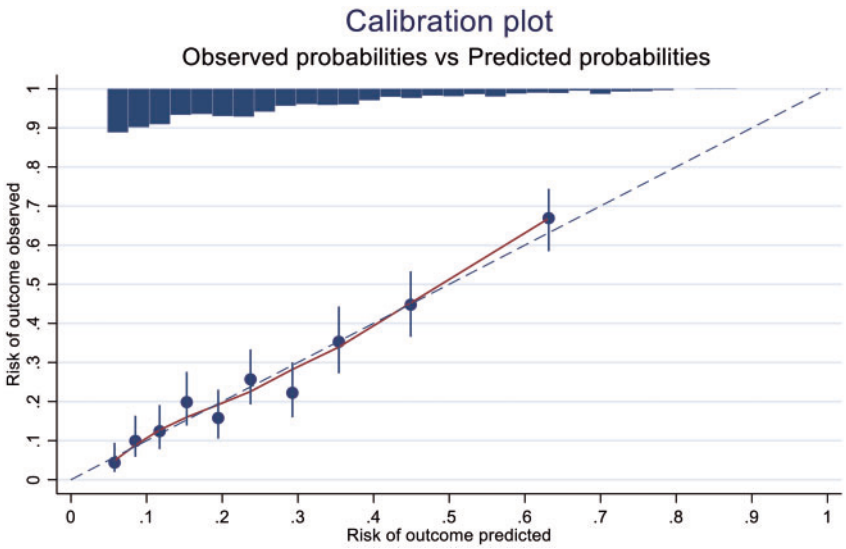


Figure 2: Calibration plot. Comparison of observed and predicted mortality rates in the cohort to assess calibration with a histogram of predicted probabilities. The sample is divided into deciles of risk. The dotted line represents the line of identity.

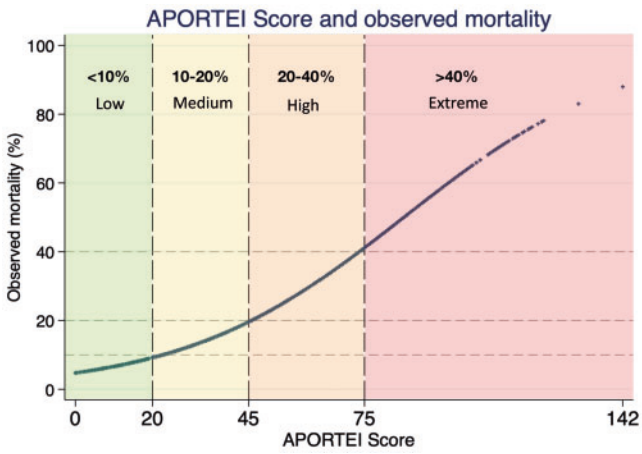


Figure 3: Scatterplot of observed deaths according to the APORTEI score. The vertical lines show the APORTEI score cut-points for which the 4 prognostic classes (low, medium, high and extreme risk) were identified. Horizontal lines show the corresponding predicted mortality rates.

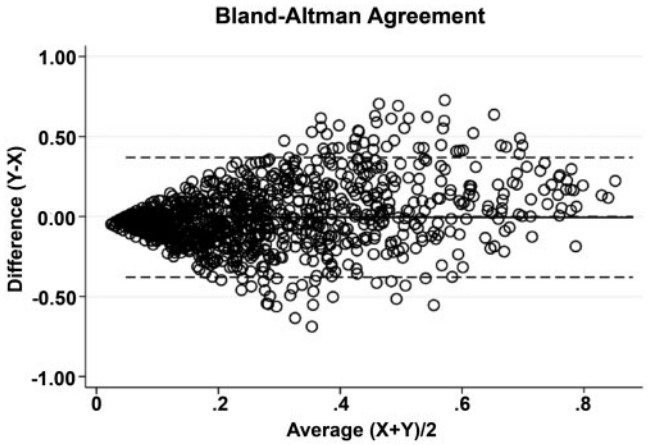


Figure 4: Bland–Altman agreement analysis graph of the predicted mortality rates of the logistic EuroSCORE I and the APORTEI score. Y represents the predicted mortality rate with the logistic EuroSCORE I and X represents the predicted mortality rate with the APORTEI score. The graph shows a clear trend that the higher the expected mortality was, the lower the agreement between both scores.

various factors such as patient baseline characteristics, preoperative status, intraoperative difficulties, postoperative care and variations in hospital expertise. However, preoperative status leads to controversies in surgical indications [10, 21]. In addition to the general factors associated with the mortality rate (e.g. pulmonary and renal disease, low ejection fraction, vascular diseases, advanced liver disease) that should be taken into consideration in establishing the indication for any cardiac operation, IESF are considered to be associated with postoperative mortality.

The univariable analysis of each of the 11 IESF [14], previously selected as mortality prognostic factors in the systematic review, showed a statistically significant relation with mortality in the GAMES cohort, the only exception being multivalvular involvement (odds ratio 1.13). Multivalvular involvement has been described as a marker of non-controlled infection, and higher mortality rates were reported in comparison with single valve involvement [22, 23]. One possible hypothesis to explain the lack of relation in our multicentric cohort is the low proportion of patients who had multivalvular involvement (only 211 patients, 15.8% of the sample), which represented a significantly lower proportion in comparison with other previously published studies that reported a multivalvular involvement of up to 26% [11, 24]. Nevertheless, in the GAMES cohort, postoperative mortality rates of patients with single and multivalvular involvement were equivalent (27.6% vs 25.22%; $P = 0.476$).

Utility of specific scores

Indications for surgery are well established in current guidelines [19]; however, it is believed that a high proportion of patients who have surgical indications do not undergo surgery [25]. A possible reason could be the estimated high mortality rate in some patients with IE, who therefore could be rejected for surgery [26]. Nevertheless, patients who have indications for surgery and are not operated on still have a dismal prognosis [27]

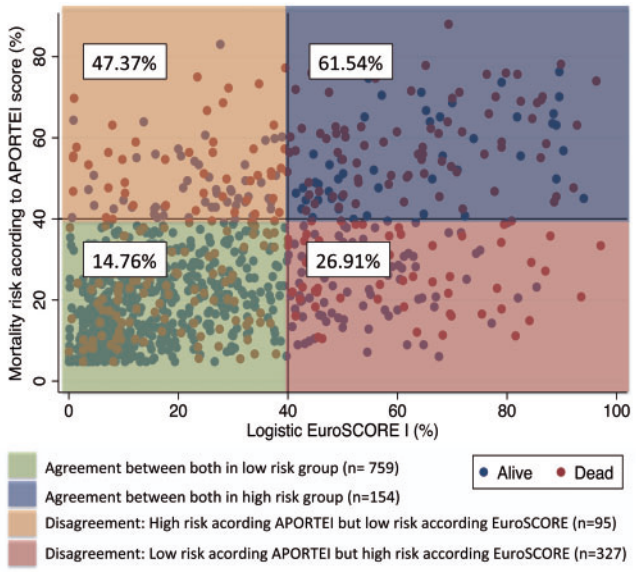


Figure 5: Scatterplot graph of the predicted mortality rate of the APORTEI score and the logistic EuroSCORE I. Y represents the predicted mortality rate with the APORTEI score and X represents the predicted mortality rate with the logistic EuroSCORE I. The graph is divided into 4 groups, classified by an estimated risk of mortality higher than 40%. Actual observed mortality in each quadrant is represented.

whereas long-term survival rates in patients who undergo surgery are acceptable [28].

It is questionable whether any predictive score may change the clinical decision process when offering an operation. Nonetheless, new specific scores have the value that they can be used to benchmark endocarditis teams and provide quality assurance in all these complex health care processes. Unfortunately, risk scores predicting mortality rates do not offer help in terms of establishing the burdens of surgical futility.

Over the last few years, a total of 12 risk scores have been postulated as possible models in preoperative mortality rate prediction in surgery for IE, including 3 classical risk scores and 9 IE-specific scores [2–9]. On the one hand, classical scores, used in daily practice, showed a suboptimal prognostic ability [21, 29]. On the other hand, specific IE scores have shown a better mortality risk prediction; however, some of them were created from relatively small data sets based on retrospective observational studies, and others are complex and difficult to calculate. Unfortunately, as a consequence, there is a lack of applicability of these scores in routine clinical practice of the multidisciplinary 'endocarditis teams'.

Previously published specific scores

The first specific risk model was developed by Costa *et al.* [7], who included patients who underwent the operation as well as patients who were treated with antibiotics. Afterwards, Gaca *et al.* [2] developed the Society of Thoracic Surgeons (STS)-IE score based on the STS database, which was validated in a large, prospective study [10]. Since its publication, the STS-IE score was one of the most used scores, and it established the basis for subsequent studies. Subsequently, the De Feo–Cotrufo score for native valve IE, the only one recommended in the European Society of Cardiology guidelines [19], showed non-inferiority compared with STS-IE [2, 29]. More recently, the EndoSCORE was published [9] after a multicentre retrospective study. Later, the AEPEI score [4] broke with the general tendency of the IE scores, and it incorporated 5 non-specific variables as independent factors of mortality.

Besides the previously mentioned scores, some of the most important scores were developed in Spain. First, López *et al.* [30] published a stratification model based on 3 variables. Martínez-Sellés *et al.* [3] developed the PALSUSE score using the GAMES national cohort. The Risk-E score [6], the first IE score that incorporates thrombocytopaenia and septic shock as IESF, was published after a multicentre prospective study. More recently, the specific EuroSCORE I and II were developed by Fernández-Hidalgo *et al.* [5] after eliminating irrelevant variables in EuroSCORE I and II and adding IESF.

Differences between the APORTEI and other specific scores

Although the previously mentioned scores were developed somewhat differently, all of them have a common aspect: development after the analysis of a cohort of patients. Developmental samples, for each score from the mentioned studies, exhibited important differences with huge discrepancies in the numbers of patients, study methods and associated diseases. The analysis of the discrimination and calibration conducted in their original studies was more than adequate, but unfortunately, some of these models performed poorly after external validation [12].

Contrarily, APORTEI is not based on a single sample of patients: it is instead the result of a systematic review of the literature. Therefore, the analysis of discrimination and validation represents an external validation of this score in a multicentric cohort of patients from 32 different Spanish hospitals. For that reason, the AUC, although adequate (AUC = 0.75), was slightly lower than those reported in the other mentioned studies when validated in their development sample. In addition, the APORTEI score showed improved calibration (HLt $P = 0.389$).

We performed a direct comparison between the APORTEI score and the logistic EuroSCORE I. We found that the EuroSCORE I underestimated the mortality rate in low-risk patients, whereas it overestimated the mortality rate in high-risk patients. The analysis of the discrepancies between both scores in our sample showed a suboptimal absolute agreement (concordance correlation coefficient = 0.55) with a clear trend: the higher the expected mortality rate, the lower was the agreement between both scores.

We believe that APORTEI requires a relatively simple and intuitive calculation, and, because of its evidence-based development from a systematic review and meta-analysis, it could be applicable to different populations.

Limitations

The major limitation of this report is the exclusion of some possible risk factors, both general risk factors and IESF (vegetation size, thrombocytopaenia, stroke or embolism).

On the one hand, some general factors that are well-known factors for mortality in cardiac surgery, such as pulmonary and renal disease, low ejection fraction, vascular diseases and liver disease, should be considered in establishing the indications for heart valve surgery. Not all of those factors are included in the APORTEI score; however, they must be taken into account when

talking about risk assessment. On the other hand, we followed prespecified strict criteria to choose the most studied variables in risk assessment by reviewing the previously published IE-specific scores. The criterion used to include 1 factor in the meta-analysis calculations was to select only those variables that were considered to be risk factors for mortality in 2 or more IE-specific scores [14]. As a result of this method, not only endocarditis-specific risk factors were included, but also some general risk factors (such as age, sex, previous surgery and functional class). However, in this manner, stroke, embolism and vegetations were excluded from the meta-analysis although they are well-known IESF that should be taken into account in predicting the risk of mortality. The addition of those variables, both general risk factors and other IESF, could have improved the estimation of mortality risk.

Secondly, a direct comparison between the APORTEI score and other specific and non-specific scores could have added more information about risk assessment and the possible benefits of the APORTEI score in routine clinical practice; however, only the logistic EuroSCORE I could be assessed with the data provided by the GAMES cohort.

Other possible limitations of the study are the risk of heterogeneity and missing data. However, the estimated percentage of missing data in the GAMES cohort is considered to be <1%, which is better than other published studies. The study included patients over a 10-year period; thus, a potential temporal effect cannot be ruled out.

CONCLUSIONS

A new specific endocarditis score (APORTEI score), developed from a systematic review and meta-analysis, showed a good estimation of the risk of mortality in the surgical treatment of active IE. The external validation of this score in a multicentric nationwide cohort confirmed adequate calibration and discrimination; therefore, it could be a useful tool for preoperative risk prediction in surgically treated patients affected with IE.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *EJCTS* online.

Conflict of interest: none declared.

Author contributions

Laura Varela Barca: Formal analysis; Investigation; Project administration; Validation; Writing—original draft. **Borja M. Fernández-Felix:** Data curation; Methodology; Validation. **Enrique Navas Elorza:** Supervision; Writing—review & editing. **Carlos A. Mestres:** Conceptualization; Formal analysis; Methodology; Writing—review & editing. **Patricia Muñoz**: Methodology; Writing—review & editing. **Gregorio Cuerpo-Caballero:** Writing—review & editing. **Hugo Rodríguez-Abella:** Writing—review & editing. **Miguel Montejo-Baranda:** Writing—review & editing. **Regino Rodríguez-Álvarez:** Writing—review & editing. **Francisco Gutiérrez Díez:** Writing—review & editing. **Miguel Angel Goenaga:** Writing—review & editing. **Eduard Quintana:** Methodology; Resources; Supervision; Writing—review & editing. **Guillermo Ojeda-Burgos:** Writing—review & editing. **Arístides de Alarcón:** Conceptualization; Methodology; Resources; Writing—review & editing. **Laura Vidal-Bonet:** Supervision; Visualization; Writing—review & editing. **Tomasa Centella Hernández:** Writing—review & editing. **Jose López-Menéndez:** Data curation; Formal analysis; Investigation; Methodology; Project administration; Supervision; Validation; Writing—original draft; Writing—review & editing.

REFERENCES

[1] Murdoch DR, Corey GR, Hoen B, Miró JM, Fowler VG, Bayer AS *et al*. Clinical presentation, etiology, and outcome of infective endocarditis in the 21st century: the International Collaboration on Endocarditis-Prospective Cohort Study. Arch Intern Med 2009;169:463–73.

[2] Gaca JG, Sheng S, Daneshmand MA, O'Brien S, Rankin JS, Brennan JM *et al*. Outcomes for endocarditis surgery in North America: a simplified risk scoring system. J Thorac Cardiovasc Surg 2011;141:98–106.

[3] Martínez-Sellés M, Muñoz P, Arnáiz A, Moreno M, Gálvez J, Rodríguez-Roda J *et al*. Valve surgery in active infective endocarditis: a simple score to predict in-hospital prognosis. Int J Cardiol 2014;175:133–7.

[4] Gatti G, Perrotti A, Obadia JF, Duval X, Iung B, Alla F *et al*.; Association for the Study and Prevention of Infective Endocarditis Study Group–Association pour l'Étude et la Prévention de l'Endocadite Infectieuse (AEPEI). Simple scoring system to predict in-hospital mortality after surgery for infective endocarditis. J Am Heart Assoc 2017;6:e004806.

[5] Fernández-Hidalgo N, Ferreira-González I, Marsal JR, Ribera A, Aznar ML, de Alarcón A *et al*. A pragmatic approach for mortality prediction after surgery in infective endocarditis: optimizing and refining EuroSCORE. Clin Microbiol Infect 2018;24:1102.

[6] Olmos C, Vilacosta I, Habib G, Maroto L, Fernández C, López J *et al*. Risk score for cardiac surgery in active left-sided infective endocarditis. Heart 2017;103:1435–42.

[7] Costa MAC, da Wollmann DR, Campos ACL, Cunha CLP, da Carvalho RG, de Andrade DF *et al*. Risk index for death by infective endocarditis: a multivariate logistic model. Rev Bras Cir Cardiovasc 2007;22:192–200.

[8] De Feo M, Cotrufo M, Carozza A, De Santo LS, Amendolara F, Giordano S *et al*. The need for a specific risk prediction system in native valve infective endocarditis surgery. Sci World J 2012;2012: 307571.

[9] Di Mauro M, Dato GMA, Barili F, Gelsomino S, Sante P, Corte AD *et al*. A predictive model for early mortality after surgical treatment of heart valve or prosthesis infective endocarditis. The EndoSCORE. Int J Cardiol 2017;241:97–102.

[10] Chu VH, Park LP, Athan E, Delahaye F, Freiberger T, Lamas C *et al*. Association between surgical indications, operative risk, and clinical outcome in infective endocarditis a prospective study from the International Collaboration on Endocarditis. Circulation 2015;131:131–40.

[11] Varela L, López-Menéndez J, Redondo A, Fajardo ER, Miguelena J, Centella T *et al*. Mortality risk prediction in infective endocarditis surgery: reliability analysis of specific scores. Eur J Cardiothorac Surg 2018;53:1049–54.

[12] Wang TKM, Wang MTM, Pemberton J. Risk scores and surgery for infective endocarditis: a meta-analysis. Int J Cardiol 2016;222:1001–2.

[13] Anantha Narayanan M, Mahfood Haddad T, Kalil AC, Kanmanthareddy A, Suri RM, Mansour G *et al*. Early versus late surgical intervention or medical management for infective endocarditis: a systematic review and meta-analysis. Heart 2016;102:950–7.

[14] Varela L, Navas E, Fernández-Hidalgo N, Moya JL, Muriel A, Fernández-Felix BM, *et al*. Prognostic factors of mortality after surgery in infective endocarditis: systematic review and meta-analysis. Infection 2019; doi: 10.1007/s15010-019-01338-x. [Epub ahead of print].

[15] Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;21:e1000097.

[16] Durack DT, Lukes AS, Bright DK. New criteria for diagnosis of infective endocarditis: utilization of specific echocardiographic findings. Duke Endocarditis Service. Am J Med 1994;96:200–9.

[17] Amilivia G. Infective endocarditis: a better outcome after surgery during the active phase. Critical Care 2000;2:6339.

[18] Akins CW, Miller DC, Turina MI, Kouchoukos NT, Blackstone EH, Grunkemeier GL *et al*. Guidelines for reporting mortality and morbidity after cardiac valve interventions. Ann Thorac Surg 2008;85:1490–5.

[19] Habib G, Lancellotti P, Antunes MJ, Bongiorni MG, Casalta J-P, Zotti FD *et al*. 2015 ESC Guidelines for the management of infective endocarditis. Eur Heart J 2015;36:3075–128.

[20] Edmunds LH, Clark RE, Cohn LH, Grunkemeier GL, Miller DC, Weisel RD. Guidelines for reporting morbidity and mortality after cardiac valvular

operations. The American Association for Thoracic Surgery, Ad Hoc Liaison Committee for Standardizing Definitions of Prosthetic Heart Valve Morbidity. Ann Thorac Surg 1996;62:932–5.
[21] Rasmussen RV, Bruun LE, Lund J, Larsen CT, Hassager C, Bruun NE. The impact of cardiac surgery in native valve infective endocarditis: can EuroSCORE guide patient selection? Int J Cardiol 2011;149:304–9.
[22] Mueller XM, Tevaearai HT, Stumpe F, Fischer AP, Hurni M, Ruchat P *et al.* Multivalvular surgery for infective endocarditis. Cardiovasc Surg 1999;7:402–8.
[23] Kim N, Lazar JM, Cunha BA, Liao W, Minnaganti V. Multi-valvular endocarditis. Clin Microbiol Infect 2000;6:207–12.
[24] Marks DJB, Hyams C, Koo CY, Pavlou M, Robbins J, Koo CS *et al.* Clinical features, microbiology and surgical outcomes of infective endocarditis: a 13-year study from a UK tertiary cardiothoracic referral centre. QJM 2015;108:219–29.
[25] Iung B, Doco-Lecompte T, Chocron S, Strady C, Delahaye F, Le Moing V *et al.* Cardiac surgery during the acute phase of infective endocarditis: discrepancies between European Society of Cardiology guidelines and practices. Eur Heart J 2016;37:840–8.
[26] Gatti G, Chocron S, Obadia J-F, Duval X, Iung B, Alla F *et al.* Using surgical risk scores in nonsurgically treated infective endocarditis patients. Hellenic J Cardiol 2019; pii: S1109-9666(18)30520-7. doi: 10.1016/j.hjc.2019.01.008. [Epub ahead of print].
[27] Ramos-Martínez A, Calderón-Parra J, Miró JM, Muñoz P, Rodríguez-Abella H, Valerio M *et al.* Effect of the type of surgical indication on mortality in patients with infective endocarditis who are rejected for surgical intervention. Int J Cardiol 2019;282:24–30.
[28] Varela Barca L, López-Menéndez J, Navas Elorza E, Moya Mur JL, Centella Hernéndez T, Redondo Palacios A *et al.* Long-term prognosis after surgery for infective endocarditis: distinction between predictors of early and late survival. Enferm Infecc Microbiol Clin 2019;37:435–40.
[29] Oh T, Voss J, Gamble G, Kang N, Pemberton J. Comparison of contemporary risk scores for predicting outcomes after surgery for active infective endocarditis. Heart Vessels 2015;30:227–34.
[30] López J, Fernandez-Hidalgo N, Revilla A, Vilacosta I, Tornos P, Almirante B *et al.* Internal and external validation of a model to predict adverse outcomes in patients with left-sided infective endocarditis. Heart 2011; 97:1138–42.

BMJ Open

Protocol for development and validation of a clinical prediction model for adverse pregnancy outcomes in women with gestational diabetes

Shamil D. Cooray ,[1,2] Jacqueline A. Boyle ,[1,3] Georgia Soldatos,[1,4] Javier Zamora ,[5,6] Borja M. Fernández Félix ,[5,7] John Allotey ,[8] Shakila Thangaratinam ,[8] Helena J. Teede [1,4]

To cite: Cooray SD, Boyle JA, Soldatos G, *et al*. Protocol for development and validation of a clinical prediction model for adverse pregnancy outcomes in women with gestational diabetes. *BMJ Open* 2020;**10**:e038845. doi:10.1136/bmjopen-2020-038845

► Prepublication history and additional material for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-038845)

ST and HJT are joint senior authors.

Received 26 March 2020
Revised 25 September 2020
Accepted 29 September 2020

For numbered affiliations see end of article.

Correspondence to
Prof Helena J. Teede;
Helena.Teede@monash.edu

ABSTRACT

Introduction Gestational diabetes (GDM) is a common yet highly heterogeneous condition. The ability to calculate the absolute risk of adverse pregnancy outcomes for an individual woman with GDM would allow preventative and therapeutic interventions to be delivered to women at high-risk, sparing women at low-risk from unnecessary care. The Prediction for Risk-Stratified care for women with GDM (PeRSonal GDM) study will develop, validate and evaluate the clinical utility of a prediction model for adverse pregnancy outcomes in women with GDM.

Methods and analysis We undertook formative research to conceptualise and design the prediction model. Informed by these findings, we will conduct a model development and validation study using a retrospective cohort design with participant data collected as part of routine clinical care across three hospitals. The study will include all pregnancies resulting in births from 1 July 2017 to 31 December 2018 coded for a diagnosis of GDM (estimated sample size 2430 pregnancies). We will use a temporal split-sample development and validation strategy. A multivariable logistic regression model will be fitted. The performance of this model will be assessed, and the validated model will also be evaluated using decision curve analysis. Finally, we will explore modes of model presentation suited to clinical use, including electronic risk calculators.

Ethics and dissemination This study was approved by the Human Research Ethics Committee of Monash Health (RES-19–0000713 L). We will disseminate results via presentations at scientific meetings and publication in peer-reviewed journals.

Trial registration details Systematic review proceeding this work was registered on PROSPERO (CRD42019115223) and the study was registered on the Australian and New Zealand Clinical Trials Registry (ACTRN12620000915954); Pre-results.

Strengths and limitations of this study

- We have designed a prediction model to meet an established clinical need by integrating learnings from a systematic review and critical appraisal of existing models, consensus from a clinical study steering committee and consideration of consumer perspectives.
- This study will build upon relevant literature, including a systematic review of existing prediction modelling studies to formulate a composite of prioritised, objective and serious adverse pregnancy outcomes and identify a broad series of relevant candidate predictors.
- We will adopt best practice methods for model development and validation framed by learnings from a critical appraisal of existing models.
- We will develop and validate the model using routinely-collected healthcare data in an ethnically and socioeconomically diverse population from multiple hospitals. This data was collected contemporaneously and prospectively, albeit not specifically for the purposes of this study hence missing data is likely.
- We will use decision curve analysis to formally evaluate the clinical utility of the model. This will inform the suitability of the validated model as a basis for risk-stratified model-of-care.

INTRODUCTION

Gestational diabetes (GDM) is diabetes that is first diagnosed during pregnancy, typically the second or third trimester of pregnancy and not consistent with pre-existing type 1 or type 2 diabetes.[1] It is a prominent health concern as it is common, affecting 7.5% to 27.0% of pregnancies,[2] and confers an increased risk of complications with health consequences for mother and baby.[3] However, current approaches to care are based on the false premise that the diagnostic criteria used define a group of women who are all at high-risk of adverse pregnancy outcomes.[4] In reality, the identified group is highly heterogeneous with a broad and continuous range of risk related to interrelated factors, which are inadequately integrated into the current glucocentric

treatment paradigm. Therefore, the ability to calculate the absolute risk of adverse pregnancy outcomes for an individual woman with GDM would support shared decision-making and a personalised approach to care. Here, the intensity of intervention could be stratified by risk of pregnancy complications such that preventative and therapeutic interventions could be delivered to women at high-risk, sparing women at low-risk from unnecessary intervention.

The International Association of Diabetes in Pregnancy Study Groups (IADPSG) diagnostic criteria sought to translate the results of the Hyperglycaemia and Adverse Pregnancy Outcome (HAPO) study into clinical practice.[4 5] This large multinational prospective cohort study demonstrated that the risk of two adverse pregnancy outcomes (birth of a large-for-gestational-age neonate, clinical neonatal hypoglycaemia), an obstetrical intervention (primary caesarean section) and a surrogate marker for fetal hyperglycaemia (cord-blood serum C-peptide >90th percentile) was positively associated with maternal glycaemia at 24 to 28 weeks gestation as measured by an oral glucose tolerance test (OGTT). The IADPSG diagnostic criteria dichotomise the risks related to GDM on serum glucose levels using an OR of 1.75 for the above outcomes. The use of an arbitrary threshold has led to disagreement among experts and professional societies.[6 7] Indeed the optimal diagnostic strategy may vary depending on the characteristics of the local population.[1 8 9] Ultimately, these diagnostic criteria have had the unintended consequence of fostering a glucocentric approach to the treatment of GDM. This study will address this need for a more refined method of risk prediction and the targeting of intervention.

The need for refined and targeted approaches is strengthened by the heterogeneous population defined by current diagnostic criteria for GDM.[10] Pregnancy risk is clearly related to elevated glucose in GDM, but the relationship is complex, and an individual's risks are modified by interrelated factors including maternal weight,[11 12] gestational weight gain,[13] ethnicity[14] and genotype.[15] For example, it has recently been shown that within the two largest maternity services in Australia, ethnic Chinese women with GDM had a lower risk of large-for-gestational-age (LGA) babies and neonatal hypoglycaemia compared with Caucasian women, even adjusting for confounders.[16] A prediction model could integrate these risk factors to estimate risk of adverse pregnancy outcome.

The feasibility of estimating an individual's absolute risk of adverse pregnancy outcomes by integrating oral glucose tolerance test results, maternal weight and pregnancy history was established in our systematic review.[17] However, critical appraisal established that existing prediction models were not yet suitable for application to clinical practice due to high risks of bias due to methodological limitations.

The Prediction for Risk-Stratified care for women with GDM (PeRSonal GDM) study will leverage the rapidly evolving methodological advances in prediction modelling to achieve the evolution required to transform promising statistical models into useful clinical tools. In this project, we integrate the findings of this systematic review and critical appraisal of existing models, pertinent findings from landmarks trials, clinical expertise and best practice methods from contemporary guidelines to inform the methodological design of the PeRSonal GDM study.

Objectives

The aims of the PeRSonal GDM study are to:

1. Develop and internally validate a prediction model for adverse pregnancy outcomes in GDM to aid shared decision-making and stratify care;
2. Externally validate the model to demonstrate temporal transportability;
3. Evaluate the clinical utility of the model as a basis for a risk-stratified model-of-care.

METHODS AND ANALYSIS

Prediction model design

We conducted formative research to conceptualise and design a robust and clinically acceptable prediction model. First, a systematic review and critical appraisal of existing prediction models for adverse pregnancy outcomes in women with GDM was conducted following a peer-reviewed protocol.[18] Second, the study steering committee comprising two obstetricians, three endocrinologists and a neonatologist formulated key clinical requirements of the prediction model (table 1). A model addressing these requirements was designed (figure 1). Finally, a multidisciplinary clinical working group was formed to provide feedback on the proposed requirements, gauge its clinical acceptability and consider its clinical application. The working group included endocrinologists (n=9), diabetes nurse educators (n=3), dieticians (n=2), midwives (n=2), administration staff (n=2) and an obstetrician (n=1) actively involved in the provision of GDM care at several maternity hospitals. We considered consumer perspectives throughout this process, from parallel qualitative research on GDM diagnosis and risk.[19]

Study design

We will conduct a prediction model development and validation study using a retrospective cohort design. It will be conducted following expert guidance for model development and validation,[20–25] and reported per the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement.[26]

Data sources and validation strategy

This study will use routinely collected health data for pregnancies resulting in a birth from 1 July 2017 to 31 December 2018 from an existing pregnancy outcomes database from a maternity service. Maternal, obstetrical

Table 1 The fundamental requirements of a prediction model for adverse pregnancy outcomes in women with gestational diabetes

Criteria	Specifications
(1) Prognostic versus diagnostic prediction model	The aim is to predict future events (prognostic prediction model)
(2) Intended scope	To inform clinicians' therapeutic decision-making and serve as a rational basis for the stratification of GDM care
(3) The target population to whom the prediction model applies	Pregnant women with GDM, per diagnostic criteria in clinical practice
(4) The outcome to be predicted	Pregnancy complications related to GDM affecting the mother (obstetrical or maternal) or the baby (fetal or neonatal)
(5) Timespan of prediction	Complications occurring during pregnancy or soon after birth
(6) Intended moment of using the model	At diagnosis of GDM, typically at 24 to 28 weeks gestation but may be earlier

Framework adapted from that originally proposed by Moons and colleagues to consider in framing a systematic review of prediction modelling studies.[48]
GDM, gestational diabetes.

and neonatal data are collected prospectively for all women booked to deliver their baby at the service. This data is collected with consent as part of routine clinical care. This data is of high-quality and completeness as it is collected under statute with the primary aim to facilitate improvements in quality of care. We will link these data deterministically to pathology data and clinical data extracted from the medical record of the parent health service. Linked pathology data is available for approximately 70% of pregnancies, and linked clinical data is available for approximately 90% of pregnancies. All collected data will be rendered non-identifiable for all research purposes, including analysis.

The data will be split by time into two groups (analysis type 2b in TRIPOD).[27] We will develop the prediction model using pregnancies resulting in births from the first 12 months of the study period (1 July 2017 to 30 June 2018). Pregnancies resulting in births from the last 6 months of the study period (1 July 2018 to 31 December 2018) will be used to evaluate the predictive performance of the developed model (external validation). This strategy will evaluate the temporal transportability of the model.

Participants

Study setting

This maternity service is one of the largest in Australia, provides universal access to healthcare comprising multiple large maternity hospitals and serves an ethnically and socioeconomically diverse population within a catchment of 1.6 million in South-East Melbourne. All

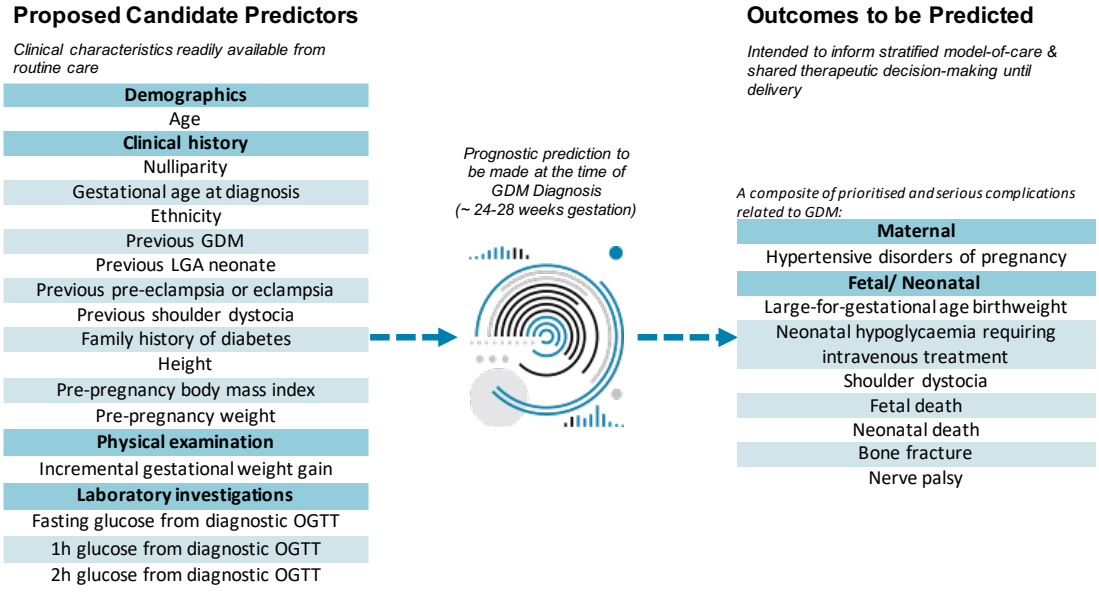


Figure 1 The design of the PeRSonal Pregnancy GDM Risk Model—Prediction for Risk-Stratified care for women with GDM. GDM, gestational diabetes; IV,intravenous; LGA, large-for-gestational-age; OGTT, oral glucose tolerance test.

levels of maternity care are available across the three hospitals with shared staff and institutional protocols and practices. Maternity care is provided to more than 9000 women each year.

Eligibility criteria

Pregnancies coded for GDM during the study period stated above will be included. There will be no exclusion criteria.

Treatment received

GDM is diagnosed and treated following institutional protocol and practices. At our service GDM is diagnosed using the International Association of Diabetes and Pregnancy Study Groups 2010 criteria,[4] as endorsed by the Australian Diabetes in Pregnancy Society with universal screening at 24 to 28 weeks with a one-step procedure using the 75 g OGTT.[6] Early screening is based on the presence of risk factors as soon as practicable using the same testing procedure with a repeat at 24 to 28 weeks if negative. The treatment package for GDM consists of an initial 2-hour group education session with diabetes nurse educator and dietician. Lifestyle management involves dietary modification, physical activity and weight management. Follow-up reviews occur with an endocrinologist or endocrinology specialist trainee every 1 to 3 weeks. Insulin is commenced where glucose targets (fasting <5.5 mmol/L and 2-hour post-prandial <7.0 mmol/L) are not met and are not amenable to further dietary modification. Metformin is used where there is evidence of significant insulin resistance, where targets are not achieved with insulin alone or when insulin use is relatively contraindicated due to the risk of significant psychological harm.

Outcome

The outcome to be predicted will be a composite consisting of a combination of eight prioritised, objective and serious adverse pregnancy outcomes defined in table 2.

Formulation of outcome(s) to be predicted

The study steering committee considered a large number of adverse pregnancy outcomes for inclusion in the composite (online supplemental table S1). Outcomes predicted by existing models identified in our systematic review and predicted by a related model for insulin therapy initiation[28] were considered. The committee also considered outcomes in the final core outcome set (COS) for GDM treatment research.[29] Reference to the COS for future GDM treatment research provided objective prioritisation of outcomes from a large international multidisciplinary group of relevant stakeholders. Finally, the committee considered all outcomes studied in the HAPO study,[5] the landmark international multicentre observational study that demonstrated associations between increasing levels of glucose levels on oral glucose tolerance testing and adverse pregnancy outcomes. From this, a composite outcome was constructed to reflect the multiple adverse pregnancy outcomes related to GDM. Construction of the composite outcome considered recommendations that components are (1) of similar importance, (2) occur with similar frequency and (3) are likely to have similar relative risk reductions (or predictive effects moving in the same direction) with similar underlying biology.[30] The rationale for inclusion or exclusion from the composite outcome to be predicted is presented in online supplemental table S2.

Table 2 The adverse pregnancy outcomes to be predicted: definition, variable type and categories

Outcome	Definition
Maternal	
Hypertensive disorders of pregnancy	Pregnancy-induced hypertension, pre-eclampsia or eclampsia
Fetal/neonatal	
LGA	Birth weight >90th percentile corrected for gestation and fetal sex using Australian population growth chart[56]
Neonatal hypoglycaemia requiring intravenous treatment	A neonate with a low blood glucose level fulfilling institutional criteria for intravenous treatment consisting of either a dextrose bolus or dextrose infusion
Shoulder dystocia	When, after delivery of the head, the baby's anterior shoulder gets caught above the mother's pubic bone
Fetal death	Death of fetus after 20 weeks gestation
Neonatal death	Death of live-born neonate
Bone fracture	Neonatal fracture (femur, humerus, clavicle or skull) suffered at birth
Nerve palsy	Neonatal nerve palsy (brachial plexus injury or facial nerve injury) suffered at birth

LGA, large-for-gestational-age.

Outcome assessment

LGA assessment will be based on a population-based growth chart rather than customised centiles to avoid incorporation of predictor information such as ethnicity into outcome assessment. Blinding to predictors in the assessment of the outcome will not be feasible.

Predictors

Definition of predictors and measurement

Candidate predictors to be evaluated for inclusion in the model are defined in table 3. There will be no blinding between the assessment of a predictor and the outcome nor to other predictors.

Identification of candidate predictors

Candidate predictors were identified from those selected for the final models included in the systematic review

Table 3 Candidate predictors to be evaluated in model development: definition, variable type and units/ categories

Candidate predictor	Definition	Variable type	Units/categories
Demographics			
Age	Mother's age	Continuous	years
Clinical history			
Nulliparity	The condition in a woman of never having given birth	Binary	0 'No' 1 'Yes'
Gestational age at diagnosis	Gestational age at diagnosis of GDM in the index pregnancy	Continuous	weeks' gestation
Ethnicity	Self-reported ethnicity with classification aligned to the Australian Standard Classification of Cultural and Ethnic Groups[57]	Categorical	Ethnicity classified into approximately five to six categories
Previous GDM	Previous diagnosis of GDM	Binary	0 'No'; 1 'Yes'
Previous LGA	Previous child with birthweight >90th percentile corrected for gestation and fetal sex using Australian population growth chart[56]	Binary	0 'No' 1 'Yes'
Previous pre-eclampsia or eclampsia	Pre-eclampsia or eclampsia in a previous pregnancy	Binary	0 'No' 1 'Yes'
Previous shoulder dystocia	Shoulder dystocia in a previous pregnancy	Binary	0 'No' 1 'Yes'
Family history of diabetes	Any family history of diabetes	Binary	0 'No' 1 'Yes'
Height	The mother's self-reported height at about the time of conception.	Continuous	centimetres (cm)
Body mass index	Body mass divided by the square of the body height	Continuous	kg/m2
Weight	Mother's self-reported weight (body mass) about the time of conception	Continuous	kilograms (kg)
Physical examination			
Incremental gestational weight gain	Weight at first GDM clinic appointment (at around 30 weeks gestation) minus preconception weight divided by gestational weeks completed at the time of the first GDM clinical appointment	Continuous	kg
Laboratory investigations			
Fasting glucose from diagnostic OGTT	Glucose level from baseline or time zero of diagnostic oral glucose tolerance test	Continuous	mmol/L
1-hour glucose from diagnostic OGTT	Glucose level 1-hour following a 75 g oral glucose load of diagnostic oral glucose tolerance test	Continuous	mmol/L
2-hour glucose from diagnostic OGTT	Glucose level 2-hour following a 75 g oral glucose load of diagnostic oral glucose tolerance test	Continuous	mmol/L

BMI, body mass index; GDM, gestational diabetes; LGA, large-for-gestational-age; OGTT, oral glucose tolerance test.

of models for pregnancy complications in women with GDM, selected in a model for GDM diagnosis previously developed by our group,[31] and selected in a related model for insulin therapy initiation.[28] (online supplemental table S3) From these existing related models 13 of the 16 predictors will be evaluated for inclusion in this prediction modelling study (table 3). Three predictors selected for related models (poor glycaemic control, enlarged abdominal circumference and HbA1c (glycatedhaemoglobin) at diagnosis) could not be evaluated in this study as the data are not routinely collected at our service.

One previous study selected history of macrosomia as a predictor for LGA.[32] Indeed, in clinical practice, past history is often seen as a major risk factor for future occurrence. Therefore, this study will evaluate previous histories of components of the composite outcome for inclusion in the model. Such data is available for macrosomia, LGA, pre-eclampsia and eclampsia, and shoulder dystocia, and therefore, these four predictors will be evaluated as candidate predictors.

In addition to the candidate predictors identified from their use in existing related models, ethnicity and gestational weight gain (GWG) were identified as potential predictors requiring formal evaluation due to the emergence of evidence supporting their role as significant prognostic factors. Chinese women affected by GDM were at a lower risk of a range of adverse pregnancy outcomes including LGA and neonatal hypoglycaemia compared with affected Caucasian women in an Australian cohort,[16] and South Asian babies exposed to GDM were smaller across gestation than babies of White European in an English cohort.[33] Emerging physiological data

suggests highly variable degrees of beta-cell function and insulin resistance among women diagnosed with GDM,[34] and that classifying women with GDM by these physiological defects may stratify women by their risk of adverse pregnancy outcomes.[35] Ethnicity may serve as a surrogate marker for these physiological defects avoiding the need for additional investigations. Hence, ethnicity is an appealing candidate predictor for models to predict the development of adverse pregnancy outcomes.

GWG has also been shown to be a risk factor for adverse pregnancy outcomes, independent of body mass index (BMI).[13] Specifically, GWG is associated with an increased proportion of LGA over and above that which is associated with GDM and overweight or obesity, in a general obstetric population.[36] BMI, parity and GWG together, better predict adverse pregnancy outcomes than BMI alone in a cohort attending a general antenatal clinic (women with GDM and normoglycaemia).[37] The effect of GWG is likely to be modified by other predictors, including ethnicity, supporting its integration within a multivariable model rather than a single prognostic factor-based approach.

Data extraction

We will extract records for eligible participants to create a research data set with each observation representing a pregnancy. Participants may be included more than once due to multiple pregnancy or repeat pregnancies within the study period. We will manually review eligible participant's medical record to ensure the accuracy of the diagnosis of GDM. Linked pathology and additional clinical data will be extracted and merged with the research data set. The research data set will be rendered non-identifiable for all subsequent analyses.

Sample size

In this study, the adequacy of the sample size of our developmental data set will be determined by the total number of events of the composite binary outcome. Approximately 9000 women are delivered annually at the institution from which the development data set will be derived. The prevalence of GDM at this institution is 18% (unpublished data). Therefore, over the 12-month period used for model development, we conservatively estimate that the development data set will include 1620 cases of women with GDM. We anticipate that at least 10% of these women will deliver neonates that have a birth weight, that is, LGA defined as greater than the 90th percentile for the population (approximately 162 events). Furthermore, using unpublished data from our institution, the prevalence of hypertensive disorders of pregnancy is 7% (approximately 113 events) and neonatal hypoglycaemia requiring intravenous treatment is 11% (approximately 178 events). Therefore the expected event count is greater than 453 once the additional contribution of the less common component outcomes are also considered (shoulder dystocia, fetal death, neonatal death, bone fracture, nerve palsy). Given we envisage including up to 20 candidate predictors, our study should be adequately powered as the data set will have in excess of 10 events per predictor as is commonly recommended to avoiding overfitting.[38]

Over the 6-month period used for external validation, the expected event count is 50% of that for the 12-month period used for development, hence approximately 225. This is greater than the recommended minimum of 100 events for validation.[39]

Missing data

We do not expect considerable missing data, but some will inevitably occur, with not all cases providing all variables of interest. Handling of missing data will be determined individually on a per predictor basis. The missing indicator method will be used for predictors where data is missing not at random. Multiple imputation by chained equations will be used to impute missing data as long as the data is missing at random. If necessary, we will include a supplementary table comparing predictor distributions between patients with missing data and patients with complete data.

Statistical analysis methods

To make individualised predictions for the binary composite of an adverse pregnancy outcome, we will apply a logistic regression model with the composite outcome as the dependent variable.

Handling of predictors

Continuous variables will be kept as continuous in the model (rather than dichotomising), to avoid a loss of prognostic information. Those predictors that are highly correlated with others contribute little information and will be excluded from the statistical analysis.

The functional form of the relationship of continuous predictors with the outcome will be assessed. If non-linear they will be modelled with fractional polynomials (FP). If this is the case, as several continuous variables were included in the model, we will use the multivariable fractional polynomial algorithm. Multiple imputation and FPs will be combined using the procedure described by Morris and colleagues.[40]

Model-bBuilding procedures (including predictor selection)

Candidate predictor variables will be selected a priori based on existing literature and clinical expertise as described above. During modelling, predictors will be selected by using a LASSO (Least Absolute Shrinkage and Selection Operator) method, which simultaneously selects the variables and penalises the model coefficients for over-optimism.[41]

Examination of predictor interactions will be undertaken for the following groups of predictors: weight, GWG and BMI, and fasting, 1-hour and 2-hour glucose levels from OGTT.

Internal validation and assessment of model performance

The model performance will be assessed in terms of discrimination and calibration. We will use a bootstrap

re-sampling technique to adjust for over-optimism in the estimation of model performance due to validation in the same data set that is used to develop the model itself. We will use the area under the curve of the receiver operating characteristic curve with 95% CI to assess the overall discriminatory ability of the developed model. We will report the apparent and adjusted for over-optimism model performance. A calibration plot will be created. This plot will facilitate the graphical assessment of calibration by putting affected women into groups ordered by predicted risk and considering the agreement between the mean predicted risk and the observed events in each risk group, usually deciles. The calibration will be summarised using the intercept and slope of the calibration plot. Internal validation, where the model's predictions are compared with the observed data, should return perfect calibration to the development data (calibration slope=1).

External validation

External validation of the developed model will be undertaken to assess temporal transportability. It will be undertaken using the model coefficients from the developed model to calculate the risk for each woman. We will report the predictive performance in a more recently treated cohort at the same maternity service using the same measures of discrimination and calibration as used in internal validation. Development and validation data are identical in terms of eligibility criteria, outcome and predictors.

Presentation of a simplified model for clinical use

Once a final model is identified, we will simplify and adapt the presentation of the model to facilitate its application to clinical practice. Alternative modes of presentation will be explored with a focus on maximising end-user usability and promoting translation into clinical care. Various presentation formats will be considered, including a simplified scoring system, nomogram and web-based or application-based electronic risk calculators.

Assessment of clinical utility

To supplement traditional measures of predictive model performance, discrimination and calibration, clinical utility will be formally evaluated. We will use decision curve analysis to explore the net benefit of developed models over the entire range of probability thresholds.[23 27 42] We will represent the net benefit as a function of the decision threshold in a decision curve plot. This will explore whether there is an overall net-benefit for using the models to stratify the population into two risk groups as a basis for a risk-stratified model of care:

1. Low-risk where the risk of adverse pregnancy outcomes is less than a pre-specified value—this group may be considered for a less intensive model-of-care;
2. High-risk where the risk is greater than a pre-specified value—this group should receive specialist-led hospital-based care.

Further formative research is planned to ascertain optimal risk thresholds. This will include engagement with stakeholders, including women affected by GDM and clinicians. A combination of focus groups and an electronic survey will be used.

Sensitivity analyses

We will conduct additional analysis to address the confounding effect of insulin treatment on predictor-outcome associations and hence the performance of the prediction model. This will consider four possible approaches with sensitivity analysis used to evaluate the robustness of each:

1. Derivation of a propensity score of being treated with insulin based on women pre-treatment characteristics. We will then weight observations by using the inverse probability of treatment weighting (IPTW). In this way, women with lower propensity to be treated will have more weight in the development of the prognostic model than those who had a higher probability of being treated.
2. Inclusion of insulin treatment as a component of the composite outcome.
3. Exclusion of cases where insulin treatment was used.
4. Exploration of the multinomial regression model framework for combinations of the composite outcome of adverse pregnancy outcome and insulin treatment.

The primary analysis will develop and validate a model based on clinical characteristics. Prognosis may also be influenced by an affected woman's capacity to implement lifestyle measures such a dietary modification and increased exercise. Therefore, we will undertake a sensitivity analysis to evaluate whether measures of socioeconomic disadvantage can improve the prediction of adverse pregnancy outcomes.

All statistical analysis will be performed using Stata V.16.1 (College Station, Texas: StataCorp LLC).

Patient and public involvement

No patient and public involvement in the development of this protocol. Patient and public perspectives will be essential to the formative research required to implement findings of this model development and validation study into clinical practice. As such patients and public will be invited to participate in this phase of our research.

DISCUSSION

Strengths

The formative research undertaken established the clinical need for a robust prediction model for adverse pregnancy outcomes in GDM to support therapeutic decision-making and stratification of care. Engagement with stakeholders in the model design stage should improve the clinical acceptability of the model and support future implementation efforts. The composite outcome of prioritised, objective and serious adverse events was formulated with reference to a systematic

review and critical appraisal of existing models (manuscript submitted for publication, 2020), the relevant core outcome set[43] and clinical expertise of endocrinologists, obstetricians and a neonatologist. This composite will be composed of LGA, neonatal hypoglycaemia, hypertensive disorders of pregnancy, shoulder dystocia, severe birth trauma (nerve palsy and bone fracture) and perinatal death. The transportability of the developed model will also be enhanced by the selection of candidate predictors using existing literature and clinical expertise, independent of the predictor-outcome association in the development data set.

Prediction of a composite outcome will more accurately quantify the multiple adverse pregnancy outcomes related to GDM and therefore, will be more translatable into clinical practice. This composite will be valid and clinically useful because the component outcomes are of similar importance, the three main components (LGA, neonatal hypoglycaemia and hypertensive disorders of pregnancy) occur with a similar frequency (approximately 10%),[44] and the predictive effects are likely to move in the same direction due to similar underlying biology.[30]

A method to estimate the absolute risk of adverse pregnancy outcomes for an individual woman affected by GDM would be of great benefit to affected woman, their clinicians and the health system. It would allow affected woman to better understand the implication of GDM on their pregnancy and facilitate shared-decision making with clinicians regarding the relative risks and benefits of interventions. At a system-level these individualised risk estimates would support a risk-stratified model-of-care which recognises the breadth and continuum of pregnancy risk attributable to GDM such that preventative and therapeutic interventions could be delivered to women at high-risk, sparing women at low-risk from low-value care. Ultimately, a robust prediction model would facilitate the transition from a glucocentric model-of-care to an individualised and holistic approach to this widespread public health problem.

Translating prediction models into clinical care is challenging.[45–47] Previous efforts of addressing this clinical prediction problem have been hampered by the use of methods, which increase the risk of biassed predictions limiting the transportability of developed models to new but related populations (manuscript submitted for publication, 2020). Thus, rigorous and robust methods have been adopted for model development and validation in this study. Methods have been framed by the learnings from our critical appraisal of existing models and will be guided by TRIPOD statement.[26]

Limitations

Use of routine-collected healthcare data

The development data set was created using routinely-collected healthcare data. This data was collected contemporaneously, and in a prospective fashion, however, they were not collected specifically for the purposes of this study. In prediction modelling studies, the use of routinely collected data enables the accruement of a greater number of events, which increases power to consider a greater number of candidate predictors without risking overfitting. However, the retrospective direction of enquiry creates the possibility of poor-quality data for both predictors and outcome, potential unmeasured predictors and as such careful evaluation of missing data and application of appropriate methods to address it are essential to minimise the effect on performance and applicability of developed models.[48]

Maternal death during pregnancy or any other complications that preclude delivery at the hospital will not be captured within the source perinatal outcomes database.

Varying diagnostic criteria

Diagnostic criteria used for GDM are controversial. Some professional societies endorse the criteria initially proposed by the International Association of Diabetes and Pregnancy Study Groups but disagreement persists.[4 6 49] There is also the acknowledgement that the optimal diagnostic strategy may vary depending on the characteristics of the local population.[1 8 9] The ideal prognostic prediction model would perform adequately across populations defined by a range of diagnostic criteria. Addressing this challenge will require developed models to be externally validated across these different populations.

Addressing treatment paradox regarding insulin use

Addressing the treatment paradox (in this case with insulin) is a challenge in prediction modelling studies. The traditional approach has been to accept predictions in the context of current care. However, this does not remove the possibility that a potentially useful model may appear to perform poorly due to the confounding effect of the judicious application of effective interventions to individual's whom clinicians subjectively assess to be at high risk of the outcome of interest.

Two solutions to address the problem of treatment paradox in prediction modelling studies have been advocated.[50] First, the use of treatments suspected to confound the predictor-outcome relationship can be set as a predictor in the final model. Second, the use of such effective treatments can be included within a composite outcome to be predicted. For this study, both approaches were considered but deemed inappropriate. For the former, the inclusion of the requirement for insulin therapy as a predictor is not possible as this information is not available at the intended moment of prediction—the time of GDM diagnosis, usually around 24 to 28 weeks gestation. For the later, inclusion of the requirement for insulin therapy within the composite outcome would impair its interpretability as this outcome occurs at a significantly higher frequency than the other component outcomes (31% vs approximately 10% based on our prior work).[44] This is likely to lead to a less meaningful composite that is primarily driven by the need for insulin therapy and no longer predicts what we want (adverse pregnancy outcomes).

While many promising novel approaches have been proposed in the statistical literature, such as multistate modelling or marginal structural models for 'treatment drop-ins',[51 52] at time of writing all are primarily based on empirical data and are yet to be applied to clinical prediction problems.

The three possible results from the sensitivity analysis to evaluate the effect of including the decision to treat with insulin will be informative and may be interpreted as follows. If the sensitivity analyses find that the inclusion of the decision to treat with insulin within the outcome:

1. Positively affects model performance, then this suggests the presence of treatment paradox, that is, pregnancy complications are more likely to occur in the absence of insulin therapy;
2. Has no significant effect on model performance then this suggests that the model is robust with predictive performance not affected by the decision to treat, that is, the absolute risk of adverse pregnancy outcomes for an individual woman with GDM is not affected by insulin therapy;
3. Negatively affects model performance, then this would suggest that adverse pregnancy outcomes are more likely to occur in women treated with insulin, and thus imply more 'severe' GDM or a harmful effect for this treatment. (unlikely)

The effect of treatment with insulin will be further evaluated using an IPTW algorithm to weight women according to their propensity of having been treated and transformation of the logistic model into a multinomial model. This multinomial model will have four categories depending on the occurrence of the composite pregnancy outcome and whether the women have received treatment with insulin or not.

The target population to whom the prediction model applies

The focus of this model and eventual clinical risk calculator is on those women who develop GDM and has been developed to address the priorities of frontline healthcare workers and services on the potential for risk stratified care for the one in five women who are diagnosed with GDM. Future work, should consider whether learnings from this project can be applied to a broader population, including pregnant women without GDM in particular those with maternal overweight or obesity.

ETHICS AND DISSEMINATION

This study has been approved by the Human Research Ethics Committee of Monash Health (RES-19–0000713L). This study will be conducted in accordance with the principles of the Declaration of Helsinki and the National Statement on Ethical Conduct in Human Research (2018).[53 54] All analyses will be conducted using non-identifiable data extracted from a pre-existing data set. The data is collected as part of routine clinical care for the primary purpose of improving the quality of pregnancy care. Consent was not obtained for the secondary use of this data because it is not practical to do so, and this research is consistent with the primary purpose for which it was collected. This study has been registered on the Australian and New Zealand Clinical Trials Registry (ACTRN12620000915954).[55] Results will be disseminated via presentation at scientific meetings and publication in peer-reviewed journals.

Author affiliations
[1]Monash Centre for Health Research and Implementation, School of Public Health and Preventative Medicine, Monash University, Clayton, Victoria, Australia
[2]Diabetes Unit, Monash Health, Clayton, Victoria, Australia
[3]Monash Women's Program, Monash Health, Clayton, Victoria, Australia
[4]Diabetes and Endocrinology Units, Monash Health, Clayton, Victoria, Australia
[5]CIBER Epidemiology and Public Health, Madrid, Comunidad de Madrid, Spain
[6]Clinical Biostatistics Unit, Hospital Ramon y Cajal, Madrid, Madrid, Spain
[7]Clinical Biostatistics Unit, Hospital Universitario Ramon y Cajal, Madrid, Madrid, Spain
[8]WHO Collaborating Centre for Global Women's Health, Institute of Metabolism and Systems Research, University of Birmingham, Birmingham, UK

Twitter Shamil D. Cooray @DrShamilCooray, Jacqueline A. Boyle @jacanab, Javier Zamora @JavierZa67, Borja M. Fernández Félix @borjamfernandez, John Allotey @JoAllotey, Shakila Thangaratinam @thangaratinam and Helena J. Teede @HelenaTeede

Acknowledgements We thank Dr Alice Stewart for providing a neonatology perspective in the study steering committee. We also thank Dr Jennifer Wong and Assistant Professor Arul Earnest for their constructive feedback throughout this project.

Contributors Conceptualisation: SDC, GS, JB, ST, HJT. Funding acquisition: SDC, JZ, ST, HJT. Investigation: SDC, JB, GS, JZ, BFF, JA, ST, HJT. Project administration: SDC, ST, HJT. Resources: SDC, ST, HJT. Supervision: JB, GS, JZ, ST, HJT. Validation: SDC, JZ, BFF, JA, ST, HJT. Visualisation: SDC, HJT. Writing – original draft: SDC, BFF, JZ, HJT. Writing – review and editing: SDC, JB, GS, JZ, BFF, JA, ST, HJT.

Funding SDC is supported by a National Health and Medical Research Council (NHMRC) Postgraduate Scholarship, a Diabetes Australia Research Program NHMRC Top-up Scholarship, the Australian Academy of Science's Douglas and Lola Douglas Scholarship and an Australian Government Department of Education and Training Endeavour Research Leadership Award. JB is supported by a Career Development Fellowship funded by the NHMRC. HJT is supported by an NHMRC Fellowship funded by the Medical Research Future Fund. BFF is supported by CIBER (Biomedical Research Network in Epidemiology and Public Health), Madrid, Spain. The funding bodies had no role in the study design, the collection, analysis and interpretation of the data, the writing of the report nor the decision to submit the paper for publication.

Competing interests SDC reports grants from the National Health and Medical Research Council (NHMRC), Diabetes Australia, the Australian Academy of Science and the Australian Government Department of Education and Training during the conduct of the study; JB reports grants from the NHMRC during the conduct of the study; BFF reports grants from CIBER (Biomedical Research Network in Epidemiology and Public Health, Madrid, Spain) during the conduct of the study and HJT reports grants from the NHMRC and the Medical Research Future Fund during the conduct of the study; no other relationships or activities that could appear to have influenced the submitted work.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Shamil D. Cooray http://orcid.org/0000-0002-6825-4440
Jacqueline A. Boyle http://orcid.org//0000-0002-3616-1637
Javier Zamora http://orcid.org/0000-0003-4901-588X
Borja M. Fernández Félix http://orcid.org/0000-0002-8798-019X
John Allotey http://orcid.org/0000-0003-4134-6246
Shakila Thangaratinam http://orcid.org/0000-0002-4254-460X
Helena J. Teede http://orcid.org/0000-0001-7609-577X

REFERENCES

1 American Diabetes Association. 2. Classification and Diagnosis of Diabetes: *Standards of Medical Care in Diabetes—2019*. *Diabetes Care* 2019;42:S13–28.
2 International Diabetes Federation. *Prevalence of gestational diabetes mellitus (GDM), % Brussels, Belgium: international diabetes Federation*. 9th edn, 2019. https://diabetesatlas.org/data/en/indicators/14/
3 Buchanan TA, Xiang AH, Page KA. Gestational diabetes mellitus: risks and management during and after pregnancy. *Nat Rev Endocrinol* 2012;8:639–49.
4 Metzger BE, Gabbe SG, Persson B, *et al*. International association of diabetes and pregnancy study groups recommendations on the diagnosis and classification of hyperglycemia in pregnancy: response to Weinert. *Diabetes Care* 2010;33:e98–82.
5 Metzger BE, Lowe LP, HAPO Study Cooperative Research Group. Hyperglycemia and adverse pregnancy outcomes. *N Engl J Med* 2008;358:1991–2002.
6 Nankervis A, McIntyre HD, Moses RG, *et al*. ADIPS consensus guidelines for the testing and diagnosis of hyperglycaemia in pregnancy in Australia and New Zealand 2014.
7 National Institute for Health and Care Excellence. Diabetes in pregnancy: management of diabetes and its complications from preconception to the postnatal period. diabetes in pregnancy: management of diabetes and its complications from preconception to the postnatal period. *London* 2015.
8 Feig DS, Berger H, Donovan L, *et al*. Diabetes and pregnancy. *Can J Diabetes* 2018;42:S255–82.
9 Committee on Practice Bulletins-Obstetrics. ACOG practice Bulletin No. 190: gestational diabetes mellitus. *Obstet Gynecol* 2018;131:e49–64.
10 Rudland VL, Wong J, Yue DK, *et al*. Gestational diabetes: seeing both the forest and the trees. *Curr Obstet Gynecol Rep* 2012;1:198–206.
11 Scifres C, Feghali M, Althouse AD, *et al*. Adverse outcomes and potential targets for intervention in gestational diabetes and obesity. *Obstet Gynecol* 2015;126:316–25.
12 Huet J, Beucher G, Rod A, *et al*. Joint impact of gestational diabetes and obesity on perinatal outcomes. *J Gynecol Obstet Hum Reprod* 2018;47:469–76.
13 Goldstein RF, Abell SK, Ranasinha S, *et al*. Association of gestational weight gain with maternal and infant outcomes: a systematic review and meta-analysis. *JAMA* 2017;317:2207–25.
14 Yuen L, Wong VW, Simmons D. Ethnic disparities in gestational diabetes. *Curr Diab Rep* 2018;18:68.
15 Hughes AE, Nodzenski M, Beaumont RN, *et al*. Fetal genotype and maternal glucose have independent and additive effects on birth weight. *Diabetes* 2018;67:1024–9.
16 Wan CS, Abell S, Aroni R, *et al*. Ethnic differences in prevalence, risk factors, and perinatal outcomes of gestational diabetes mellitus: a comparison between immigrant ethnic Chinese women and Australian-born Caucasian women in Australia. *J Diabetes* 2019;11:809–17.
17 Cooray SD, Wijeyaratne LA, Soldatos G, *et al*. The unrealised potential for predicting pregnancy complications in women with gestational diabetes: a systematic review and critical appraisal. *Int J Environ Res Public Health* 2020;17:3048.
18 Cooray SD, Boyle JA, Soldatos G, *et al*. Prognostic prediction models for pregnancy complications in women with gestational diabetes: a protocol for systematic review, critical appraisal and meta-analysis. *Syst Rev* 2019;8:270.
19 Wan CS, Nankervis A, Teede H, *et al*. Ethnicity and gestational diabetes mellitus care: providers' and patients' perspectives. *Qual Health Res*2020.
20 Royston P, Moons KGM, Altman DG, *et al*. Prognosis and prognostic research: developing a prognostic model. *BMJ* 2009;338:b604.
21 Altman DG, Vergouwe Y, Royston P, *et al*. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;338:b605.
22 Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*. Second edition. New York, London: Springer International Publishing, 2019.
23 Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;35:1925–31.
24 Wynants L, Collins GS, Van Calster B. Key steps and common pitfalls in developing and validating risk models. *BJOG: Int J Obstet Gy* 2017;124:423–32.
25 Steyerberg EW, Moons KGM, van der Windt DA, *et al*. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10:e1001381.
26 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162:55–63.
27 Moons KGM, Altman DG, Reitsma JB, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1–73.
28 Barnes RA, Wong T, Ross GP, *et al*. A novel validated model for the prediction of insulin therapy initiation and adverse perinatal outcomes in women with gestational diabetes mellitus. *Diabetologia* 2016;59:2331–8.
29 Egan AM, Bogdanet D, Griffin TP, *et al*. A core outcome set for studies of gestational diabetes mellitus prevention and treatment. *Diabetologia* 2020;63:1120–7.
30 Montori VM, Permanyer-Miralda G, Ferreira-González I, *et al*. Validity of composite end points in clinical trials. *BMJ* 2005;330:594–6.
31 Teede HJ, Harrison CL, Teh WT, *et al*. Gestational diabetes: development of an early risk prediction tool to facilitate opportunities for prevention. *Aust N Z J Obstet Gynaecol* 2011;51:499–504.
32 Tomlinson TM, Mostello DJ, Lim K-H, *et al*. Fetal overgrowth in pregnancies complicated by diabetes: development of a clinical prediction index. *Arch Gynecol Obstet* 2018;298:67–74.
33 Brand JS, West J, Tuffnell D, *et al*. Gestational diabetes and ultrasound-assessed fetal growth in South Asian and white European women: findings from a prospective pregnancy cohort. *BMC Med* 2018;16:203.
34 Powe CE, Allard C, Battista M-C, *et al*. Heterogeneous contribution of insulin sensitivity and secretion defects to gestational diabetes mellitus: table 1. *Diabetes Care* 2016;39:1052–5.
35 Benhalima K, Van Crombrugge P, Moyson C, *et al*. Characteristics and pregnancy outcomes across gestational diabetes mellitus subtypes based on insulin resistance. *Diabetologia* 2019;62:2118–28.
36 Black MH, Sacks DA, Xiang AH, *et al*. The relative contribution of prepregnancy overweight and obesity, gestational weight gain, and IADPSG-defined gestational diabetes mellitus to fetal overgrowth. *Diabetes Care* 2013;36:56–62.
37 Magann EF, Doherty DA, Chauhan SP, *et al*. Pregnancy, obesity, gestational weight gain, and parity as predictors of Peripartum complications. *Arch Gynecol Obstet* 2011;284:827–36.
38 Moons KGM, Wolff RF, Riley RD, *et al*. Probast: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med* 2019;170:W1–33.
39 Vergouwe Y, Steyerberg EW, Eijkemans MJC, *et al*. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005;58:475–83.
40 Morris TP, White IR, Carpenter JR, *et al*. Combining fractional polynomial model building with multiple imputation. *Stat Med* 2015;34:3298–317.
41 Tibshirani R. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B* 1996;58:267–88.
42 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;26:565–74.
43 Egan AM, Dunne FP, Biesty LM, *et al*. Gestational diabetes prevention and treatment: a protocol for developing core outcome sets. *BMJ Open*2019;9:e030574.
44 Abell SK, Boyle JA, Earnest A, *et al*. Impact of different glycaemic treatment targets on pregnancy outcomes in gestational diabetes. *Diabet. Med.* 2019;36:177–83.
45 Wyatt JC, Altman DG. Commentary: prognostic models: clinically useful or quickly forgotten? *BMJ* 1995;311:1539–41.

46 Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* 2006;144:201–9.
47 Kleinrouweler CE, Cheong-See FM, Collins GS, *et al*. Prognostic models in obstetrics: available, but far from applicable. *Am J Obstet Gynecol* 2016;214:e36:79–90.
48 Moons KGM, de Groot JAH, Bouwmeester W, *et al*. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the charms checklist. *PLoS Med* 2014;11:e1001744.
49 National Institute for Health and Care Excellence. Diabetes in pregnancy: management from preconception to the postnatal period (NICE guideline [NG3]). London; 2015.
50 Cheong-See F, Allotey J, Marlin N, *et al*. Prediction models in obstetrics: understanding the treatment paradox and potential solutions to the threat it poses. *BJOG: Int J Obstet Gy* 2016;123:1060–4.
51 Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med* 2007;26:2389–430.
52 Sperrin M, Martin GP, Pate A, *et al*. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. *Stat Med* 2018;37:4142–54.
53 The World Medical Association. WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects. 18th WMA General Assembly. Helsinki; 1964.
54 The National Health and Medical Research Council, The Australian Research Council, Universities Australia. National statement on ethical conduct in human research 2007 (updated 2018). Canberra Commonwealth of Australia; 2007.
55 Australian and New Zealand Clinical Trials Registry. The prediction modelling for risk-stratified care for women with gestational diabetes (personal GDM) study: calculating the individualised risk of adverse outcomes for women with gestational diabetes (ACTRN12620000915954) Sydney, Australia: NHMRC clinical trials centre, University of Sydney, 2020. Available: https://www.anzctr.org.au/ACTRN12620000915954.aspx [Accessed 25 Sep 2020].
56 Dobbins TA, Sullivan EA, Roberts CL, *et al*. Australian National birthweight percentiles by sex and gestational age, 1998–2007. *Med J Aust* 2012;197:291–4.
57 Australian Bureau of Statistics. 1249.0 - Australian Standard Classification of Cultural and Ethnic Groups (ASCCEG). Canberra Commonwealth Government; 2016. https://www.abs.gov.au/ausstats/abs@.nsf/mf/1249.0

CAPÍTULO 15

c0075 Lectura crítica de revisiones sistemáticas de estudios de pronóstico o riesgo

Miguel Maldonado Fernández ■ Borja Manuel Fernández Félix ■ Juan Bautista Cabello López

OBJETIVOS DEL CAPÍTULO

p0360 u0365 u0370

- Definir los distintos tipos de estudios pronósticos.
- Revisar el formato de pregunta clínica empleado específicamente en este tipo de revisiones.
- Comentar los pasos que dan los autores de estas revisiones, haciendo hincapié en los puntos más importantes para la lectura crítica de este tipo de revisiones.

st0015 Introducción

p0385 Hacer un pronóstico consiste en conocer el futuro. Los estudios de pronóstico buscan averiguar qué le sucederá a un paciente afectado por una determinada circunstancia como una enfermedad, un factor de riesgo o un tratamiento. Este tipo de estudios son especialmente importantes para los pacientes y sus familiares. No obstante, los estudios de pronóstico son especialmente complejos y difíciles de llevar a cabo. La lectura crítica de revisiones sistemáticas de estudios pronósticos posee unas peculiaridades respecto a otro tipo de revisiones sistemáticas. Se deberá tener en cuenta el tipo de estudio de pronóstico analizado. La pregunta clínica tiene un formato PICO-TA. Se emplean herramientas específicas para la extracción de datos de los estudios individuales (CHARMS-PF) y para el estudio del riesgo de sesgo (QUIPS). La certidumbre en la evidencia (calidad de la evidencia) depende del riesgo de sesgo en cada estudio individual y, además, del riesgo de sesgo de la propia revisión por factores como la imprecisión, el sesgo de publicación o la existencia de evidencia indirecta, entre otros. GRADE es una herramienta que mide la certeza en la evidencia. La heterogeneidad entre estudios pronósticos es frecuente, por lo que se recomienda usar un modelo de efectos aleatorios para el metaanálisis y presentar intervalos de predicción para la estimación del efecto. Las revisiones de factores pronóstico son a menudo complejas, pues suelen presentar dificultades tales como sesgo de publicación o reporte selectivo, diferencias en la elección de los puntos de corte o distintos factores de ajuste.

st0020 Escenario

p0390 Formas parte de la comisión de guías y protocolos de tu hospital y en las últimas reuniones se ha planteado como objetivo actualizar los protocolos generales de reanimación cardiopulmonar (RCP) que llevan algunos años sin actualizar. En realidad piensan que será un proceso complejo con múltiples estratos que incluye desde la revisión de la infraestructura institucional, los protocolos (pre-, intra- y posparada), la actualización del sistema de recogida de datos en coherencia con los

ISBN: 978-84-9113-883-9; PII: B978-84-9113-883-9.00015-0; Autor: CABELLOSPAIN2021; Documento ID: 00015; Capítulo ID: c0075

estándares internacionales para registros de investigación sobre RCP (GWTG-R registry, UK National Cardiac Arrest. The Utstein register: DOI: 10.1161/CIR.0000000000000710) y también la promoción de campañas para generar reflexiones y cambios culturales al respecto en el hospital y en el área de salud.

p0395 En todo caso, el grupo promotor de la comisión y los documentalistas han hecho una búsqueda muy amplia y parece que hay muchos frentes que revisar. Entre los documentos que destacan está una clásica recomendación de expertos de la American Heart Association (DOI: 10.1161/CIR.0b013e31828b2770), y también aparecen bastantes estudios de factores pronósticos y una interesante revisión sistemática que resume y sintetiza todos los estudios de ese tipo.

p0400 Fernando SM, Tran A, Wei Cheng, et al. Pre-arrest and intra-arrest prognostic factors associated with survival after in-hospital cardiac arrest: systematic review and meta-analysis. BMJ. 2019;367:l6373.

p0405 Como saben de tu interés y de tus habilidades en la lectura critica de estudios, te encargan revisar en profundidad ese artículo y comentarlo en la próxima sesión de la comisión.

p0410 De modo que te tocara opinar sobre:

o0010 **1.** ¿Qué factores influyen en el pronóstico de las paradas intrahospitalarias?

o0015 **2.** ¿Algunos de esos factores o grupos de factores deberían ser considerados en la elaboración de la nueva estrategia?

p0425 Lee el artículo y contesta a esas preguntas.

st0025 # Puntos clave para la lectura de una revisión sistemática de estudios pronósticos

st0030 ## PREGUNTA PICO-TA

p0430 Para definir adecuadamente la pregunta que busca responder la revisión sistemática es necesario establecer el tipo de estudios que nos interesan según la clasificación PROGRESS. Recordemos que, en este capítulo, hablaremos específicamente del PROGRESS tipo II. Los estudios de cohortes prospectivas son los idóneos para llevar a cabo estas revisiones. No obstante, en ocasiones no se dispone de cohortes prospectivas sino de otro tipo de estudios, como cohortes históricas.

p0435 En las revisiones sistemáticas de estudios de tratamiento se utiliza el formato PICO (Paciente, Intervención, Control y *Outcome* o resultado) para construir la pregunta clínica. Para las revisiones de estudios pronóstico se adapta la pregunta PICO al formato PICO-TS, en español PICO-TA, que es el acrónimo de:

u0375 **Población:** población general en la que se estudiará el factor pronóstico.

u0380 **Índice:** factor pronóstico que se está analizando.

u0385 **Comparador(es)/control(es):** define dos conceptos: el comparador, otro factor de riesgo con el que se desea comparar el índice pronóstico bajo revisión; o el confusor (cuando el propósito no es comparativo) que es un factor de ajuste considerado en la estimación del efecto del factor pronóstico bajo revisión.

u0390 ***Outcome:*** resultado o evento que se está intentando predecir (por ejemplo, mortalidad por infarto agudo de miocardio).

u0395 **Tiempo *(Timing):*** cuándo se mide el factor pronóstico y en qué lapso de tiempo se predecirá el *outcome* o desenlace.

u0400 **Ámbito *(Setting):*** el escenario donde se utilizará el factor pronóstico.

st0035 ## ESTRATEGIA DE BÚSQUEDA

p0405 La búsqueda de los estudios individuales es más compleja en las revisiones pronósticas por el hecho de que los estudios no suelen estar etiquetados como «pronósticos» y, por lo tanto, existe el riesgo de que no se detecten mediante una estrategia de búsqueda convencional. Otra dificultad añadida

ISBN: 978-84-9113-883-9; PII: B978-84-9113-883-9.00015-0; Autor: CABELLOSPAIN2021; Documento ID: 00015; Capítulo ID: c0075

es que no existen filtros metodológicos de búsqueda que hayan sido validados. Se han desarrollado y validado filtros metodológicos para la identificación de estudios de modelos pronóstico (7), que han mostrado relativa capacidad para identificar estudios de factores pronóstico.

p0475 Por lo tanto, por miedo a perder estudios primarios importantes, suele realizarse una estrategia más amplia, con el inconveniente de que se obtienen muchos resultados que no son relevantes. Cuando la pregunta de revisión se centra en un factor pronóstico, desenlace o población específica, añadir estos términos en la estrategia de búsqueda reducirá considerablemente el número de artículos identificados.

st0040 ### Sesgo de publicación y reporte selectivo

p0480 Uno de los problemas más importantes que nos encontramos en las revisiones sistemáticas de estudios de factores pronóstico es el sesgo de publicación. Bien es sabido que las revistas científicas son más proclives a publicar estudios que presentan resultados, digamos, estadísticamente «significativos». Pero no menos importante es que los investigadores no envíen a publicar aquellos estudios en los que no se encontraron los resultados esperados, y cuando estos se encontraron no es poco frecuente que sea tras buscar y escarbar sobremanera en los datos, dando lugar al reporte selectivo. Todo ello pone de manifiesto la importancia de los protocolos en la investigación del pronóstico.

st0045 ## PROTOCOLO

p0485 Al leer una revisión sistemática de estudios pronósticos se debería considerar si los autores han cumplido con sus objetivos iniciales. Si no lo han hecho, se podría estar incurriendo en algún tipo de sesgo. El lector debería poder consultar dicha información en el protocolo. Este documento es un plan detallado de lo que se va a hacer y cómo se va a llevar a cabo en la revisión sistemática. El protocolo incluye el razonamiento y la justificación de la revisión (¿por qué hay que hacer esta revisión?); los objetivos; los criterios de elegibilidad de los estudios; el método de extracción de datos; la evaluación crítica; los métodos estadísticos para sintetizar la «evidencia» (el resultado global de la revisión); y la redacción *(report)* clara y completa de los resultados. Se puede buscar el protocolo en PROSPERO, un registro específico que depende de la Universidad de York (https://www.crd.york.ac.uk/prospero/), o publicado en ciertas revistas científicas. Se trata, en cualquier caso, de ofrecer al lector de la revisión la opción de contrastar los objetivos iniciales de los autores con lo que finalmente se presentó.

st0050 ## CÓMO SE HA EXTRAÍDO LA INFORMACIÓN DE CADA ESTUDIO

p0490 La herramienta CHARMS (CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) es una lista con los ítems que deben extraerse de los estudios individuales que se incluyen en una revisión sistemática de estudios pronósticos (8). Aunque CHARMS está diseñada para las revisiones de modelos pronósticos (es decir, los PROGRESS tipo III), existe una adaptación para las revisiones de estudios de factores pronósticos, que recibe el imaginativo nombre de CHARMS-PF (6). Los ítems incluidos vienen recogidos en la tabla 15.1.

st0055 ## RIESGO DE SESGO

p0495 El siguiente paso en la revisión es comprobar cuán creíble es la información que hemos encontrado, es decir, cuál es el riesgo de sesgo de cada uno de los estudios. Es probable encontrarnos en algunas publicaciones «riesgo de sesgo». Para medir el riesgo de sesgo existe la herramienta QUIPS, acrónimo de QUality In Prognosis Studies (9), que evalúa las siguientes «dominios» (fig. 15.1):

u0405
- **Participación en el estudio.** En este ítem se comprueba si la relación entre el factor pronóstico y el desenlace puede ser diferente en los participantes en el estudio que estamos analizando, comparado con la «población elegible» que no está en ese estudio. Por eso se describe la fuente

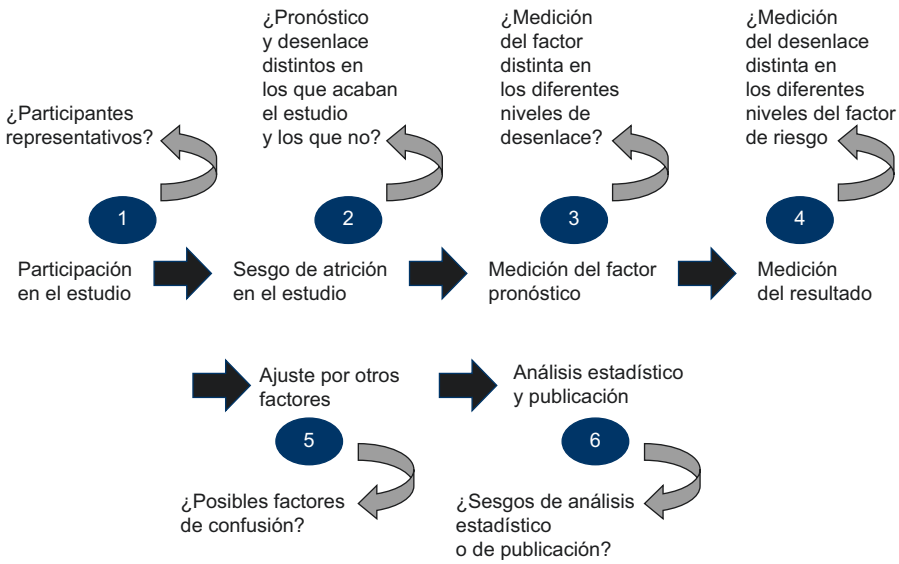
ISBN: 978-84-9113-883-9; PII: B978-84-9113-883-9.00015-0; Autor: CABELLOSPAIN2021; Documento ID: 00015; Capítulo ID: c0075

t0010 TABLA 15.1 ■ **Dominios analizados en CHARMS-PF**

Área o dominio	Comentarios	
Origen de los datos	• (Ensayo clínico, estudio de cohortes, casos y controles, etc.)	u0010
Participantes	• Elegibilidad de los participantes y método de selección	u0015
	• Descripción de los participantes	u0020
	• Detalles de los tratamientos recibidos, si fuere relevante	u0025
	• Fechas de los estudios	u0030
Desenlaces *(outcomes)* que serán medidos	• Definición del desenlace y método para su medición	u0035
	• ¿Se han usado la misma definición y método de medición en todos los participantes?	u0040
	• ¿Desenlace único o combinado?	u0045
	• ¿Hubo enmascaramiento para el desenlace?	u0050
	• ¿Tiempo hasta la aparición del desenlace?	u0055
Predictores candidatos	• Número y tipo de predictores (por ejemplo: características demográficas, historia del paciente, exploración física, nuevas pruebas diagnósticas, características de la enfermedad)	u0060
	• Definición del predictor candidato y método para su medición	u0065
	• Momento de la medición del factor predictor (o pronóstico)	u0070
	• ¿Se enmascararon de los factores para el resultado? ¿Se enmascararon entre sí?	u0075
	• ¿Cómo se tratan los factores en el modelo de predicción? (variable continua, categórica, transformación lineal, transformación no lineal)	u0080
Tamaño de la muestra	• ¿Se calculó el tamaño de la muestra? ¿Cómo?	u0085
	• Número de participantes y de eventos	u0090
	• Número de desenlaces/eventos por cada factor pronóstico	u0095
Datos que faltan	• Número de participantes en los que falta algún valor (incluyendo factores pronósticos y desenlaces)	u0100
	• Número de participantes en los que falta algún dato para cada uno de los factores pronósticos	u0105
	• Datos de atrición. En estudios de supervivencia, número de observaciones censuradas	u0110
	• Cómo se manejaron los datos que faltan (imputación de datos faltantes, análisis de casos completos, etc.)	u0115
Análisis	• Tipo de modelo (Logístico, lineal, Cox, etc.)	u0120
	• Cómo se comprobaron las asunciones del modelo	u0125
	• Método empleado para la selección de factores candidatos en el modelo multivariante	u0130
	• Método de selección de factores durante el modelado multivariante (selección retrógrada o anterógrada) y criterios para la selección (valor de la *p*; Criterio de información de Akaike)	u0135
	• Métodos para el manejo de factores continuos (dicotomización, categorización, lineal, no lineal), incluyendo los puntos de corte elegidos y su justificación	u0140
Resultados	• Estimaciones del efecto pronóstico, crudas y ajustadas, junto con sus intervalos de confianza correspondientes	u0145
	• Para cada estimación ajustada, indicar para qué factor se ajustó	u0150
Interpretación	• Interpretación de los resultados presentados	u0155
	• Comparación con otros estudios	u0160

Modificado de Riley 2019 (6).

ISBN: 978-84-9113-883-9; PII: B978-84-9113-883-9.00015-0; Autor: CABELLOSPAIN2021; Documento ID: 00015; Capítulo ID: c0075



f0010 **Figura 15.1** Dominios analizados por QUIPS. (Modificado de Hayden 2013 [9]).

de la muestra extraída, las características basales de esa muestra, cómo se ha obtenido y durante cuánto tiempo, y cuáles son los criterios de inclusión y exclusión. En resumen: los participantes del estudio ¿son «similares» a la población donde usaremos los factores pronósticos?

u0410 ■ **Sesgo de atrición en el estudio.** Analizaremos si la relación entre factor pronóstico y desenlace es probablemente muy distinta entre los que han concluido el estudio y aquellos que no han llegado al final (o «se han perdido» por el camino).

u0415 ■ **Medición del factor pronóstico.** Investigaremos si la medición del factor pronóstico se ha llevado a cabo de manera diferente en los distintos niveles del desenlace *(outcome)* estudiado. Si se midiese de modo más exhaustivo el factor pronóstico en los que han desarrollado el desenlace comparado con los que no lo han desarrollado, podríamos encontrar que el factor se relaciona con el desenlace pero que este hallazgo no sea cierto.

u0420 ■ **Medición del resultado.** La medición del resultado o desenlace *(outcome)* ¿se ha realizado de modo distinto en los diferentes niveles de factor pronóstico (es decir, más en los que han estado expuestos al factor que entre los que no han estado expuestos, por ejemplo)?

u0425 ■ **Ajuste por otros factores.** Comprobaremos si la relación entre el factor pronóstico y el desenlace es probable que esté influida por otra variable o factor que esté relacionado con el factor pronóstico estudiado y con el desenlace. Es decir, estudiaremos si es posible que existan factores de confusión.

u0430 ■ **Análisis estadístico y publicación.** Consideraremos si el resultado (del estudio individual que estamos analizando) sea espurio y realmente se deba a sesgos en el análisis estadístico o en la publicación de resultados.

st0060 ¿SE HA HECHO UN METAANÁLISIS?

p0530 El metaanálisis no siempre es el producto final de la revisión sistemática. Un metaanálisis solo se debe llevar a cabo cuando los estudios identificados sean suficientemente robustos y comparables,

ISBN: 978-84-9113-883-9; PII: B978-84-9113-883-9.00015-0; Autor: CABELLOSPAIN2021; Documento ID: 00015; Capítulo ID: c0075

de modo que los resultados derivados de este tengan una interpretación y un impacto directo en los cuidados de salud. El metaanálisis requiere de, al menos, dos estudios que estimen el mismo parámetro. Cuando se agregan estudios de baja calidad, la evidencia también será de baja calidad.

p0535 Aunque nos vamos a centrar en el metaanálisis de datos agregados, es decir, de estimaciones extraídas desde los estudios identificados en la búsqueda, una alternativa es el metaanálisis de datos individuales de pacientes, el cual precisa de la información individual de los pacientes incluidos en los estudios.

p0540 Para el metaanálisis de datos agregados son varios los obstáculos que se suelen encontrar, y que dificultan la interpretación de los resultados. Algunas de las dificultades más comunes en los estudios de factores pronóstico son: *primero,* diferentes tipos de estimaciones, en este tipo de estudios es frecuente encontrar la estimación del tamaño del efecto en términos de riesgo relativo (RR), *odds ratio* (OR) o *hazard ratio* (HR) cuando la variable de resultado es dicotómica, o la diferencia media cuando la variable es cuantitativa; *segundo,* estimaciones sin errores estándar, cuando en los métodos de metaanálisis estándar se emplean para ponderar el peso de cada estudio; *tercero,* estimaciones en diferentes tiempos de predicción: en ocasiones, los tiempos de predicción de un mismo desenlace o el momento de medición del factor pronóstico bajo revisión difieren entre estudios; *cuarto,* diferentes métodos o instrumentos de medida tanto para el desenlace como para el factor pronóstico; *quinto,* mezcla de estudios en los cuales la estimación del efecto fue ajustado en unos y crudo (o sin ajuste) en otros, y aunque todos los estudios reporten una estimación del efecto ajustada el conjunto de variables de ajuste (confusores) frecuentemente divergirá entre ellos. En este sentido, suele ayudar predefinir un conjunto básico de factores de ajuste que represente un ajuste mínimo necesario para la inclusión del estudio, por ejemplo que el efecto del factor pronóstico este ajustado, al menos, por género y edad; *sexto,* el manejo de factores pronóstico de tipo continuo: algunos estudios podrían considerar una relación lineal, otros ajustar tendencias no lineales, y otros establecer diferentes puntos de corte para categorizar o dicotomizar el factor pronóstico bajo revisión. Cuando el punto de corte elegido es seleccionado porque minimiza el *p*-valor asociado con el efecto pronóstico de interés, tenderá a sesgar los resultados hacia un mayor efecto pronóstico. Cuando se categoriza o dicotomiza el factor pronóstico o desenlace de interés la decisión de los puntos de corte debe establecerse *a priori.* Muchos de estos aspectos dirigen a sustancial heterogeneidad, causando que la estimación del efecto pronóstico varíe entre estudios.

p0545 Por otra parte, en el caso de realizar el metaanálisis, dada la heterogeneidad inherente de los estudios de factores pronóstico, es recomendado el uso de métodos de efectos aleatorios para considerar la heterogeneidad no explicada entre estudios.

p0550 Un metaanálisis de efectos aleatorios combina las estimaciones del efecto del factor pronóstico bajo revisión entre los estudios, obteniendo un efecto promedio (μ) y la desviación estándar a través de los estudios (τ). Si Y_i y var(Y_i) denotan la estimación del efecto y su varianza en el estudio i, en términos generales un modelo de metaanálisis de efectos aleatorios se puede especificar como:

p0555 $$Y_i \sim N(\mu, \text{var}(Yi) + \tau^2)$$

p0560 Esta simpática fórmula indica que, según el modelo de efectos aleatorios, se tienen en cuenta dos fuentes de variabilidad: la propia *dentro* de cada estudio y la variabilidad *entre* estudios.

p0565 Común a otros tipos de revisiones, existen diferentes métodos para estimar el modelo. El modelo de efectos fijos utiliza habitualmente el método de Mantel-Haenszel, mientras que para el modelo de efectos aleatorios el más frecuentemente utilizado por los investigadores es el método de DerSimonian y Laird (10), usual en metaanálisis de ensayos clínicos y de especial utilidad en el caso de los estudios de factores pronósticos por la heterogeneidad arriba señalada. En el contexto de heterogeneidad de los estudios de factores pronóstico se recomienda el método de Hartung-Knapp (11) —y, cuando el número de estudios que se van a combinar es pequeño, el método de Hartung-Knapp-Sidik-Jonkman (12)—, dado que ha demostrado ser más robusto.

ISBN: 978-84-9113-883-9; PII: B978-84-9113-883-9.00015-0; Autor: CABELLOSPAIN2021; Documento ID: 00015; Capítulo ID: c0075

p0570 Para llevar a cabo el metaanálisis se recomienda emplear la escala original solo cuando el estadístico estimado es la diferencia media (desenlace de tipo cuantitativo); cuando el estadístico que se desea agregar es un RR, OR o HR, la escala apropiada para el metaanálisis es el logaritmo neperiano. En este caso, el estadístico promedio y sus intervalos de confianza son estimados en escala logarítmica y, posteriormente, se deben transformar de nuevo a la escala original.

st0065 ## HETEROGENEIDAD

p0575 Cuando la heterogeneidad entre los estudios identificados en la revisión sistemática es substancial, la estimación promedio resultante del metaanálisis es difícil de trasladar a la práctica clínica. En tales situaciones el hallazgo principal de la revisión es la propia heterogeneidad identificada entre los estudios y la necesidad de investigar las posibles causas. La variabilidad entre los estudios puede ser mostrada mediante un *forest plot,* preferiblemente sin el resultado de la estimación promedio del efecto pronóstico del factor bajo revisión.

p0580 La heterogeneidad, como ya se ha mencionado en el capítulo 12, se puede cuantificar mediante el estadístico I^2, el cual mide el porcentaje de la variabilidad total debida a las diferencias entre estudios, y cuyo rango de valores oscila entre 0 y 100%. Valores próximos a 0% indican poca heterogeneidad, y a medida que aumenta, se incrementa la sospecha de heterogeneidad.

p0585 Si el metaanálisis se realiza a pesar de la presencia de heterogeneidad, es recomendable presentar el intervalo de predicción de la estimación del efecto pronóstico. Dicho intervalo de predicción indica el potencial valor del verdadero efecto pronóstico del factor en una nueva población a partir de los resultados de la revisión. Técnicas bayesianas también pueden ser empleadas para obtener inferencias predictivas. Por ejemplo, tras el metaanálisis se podría obtener la probabilidad de que el verdadero efecto pronóstico del factor sea superior a un valor dado (p. ej., un HR > 1,3 para un factor binario, que indica un incremento del riesgo de al menos un 30%).

p0590 Como en revisiones sistemáticas de intervenciones, análisis de subgrupos y metarregresión pueden ser empleados para explorar y examinar las potenciales causas de heterogeneidad.

st0070 ## ANÁLISIS DE SENSIBILIDAD

p0595 En ocasiones es preciso hacer un análisis a parte de algún subgrupo de estudios, que por algún motivo nos interesan de forma especial. Por ejemplo, podemos querer hacer un análisis de los estudios con poco riesgo de sesgo, excluyendo los que tienen un riesgo de sesgo elevado. O podemos querer analizar estudios realizados exclusivamente en ancianos o en personas con o sin una determinada característica, para comprobar qué resultado arroja el metaanálisis en ese caso.

st0075 ## GRADE (CERTIDUMBRE DE LA EVIDENCIA)

p0600 GRADE es el acrónimo de Grading of Recommendations Assessment, Development and Evaluation (13). Es una herramienta que mide, por un lado, la fuerza de la evidencia científica y, por otro, la fuerza de la recomendación basada en esa evidencia. La calidad de la evidencia o certidumbre depende de dos factores. Por una parte, el riesgo de sesgo en cada uno de los estudios incluidos en la revisión. Por otro, el riesgo de sesgo de la propia revisión, debido a factores como la imprecisión, los sesgos de publicación, la evidencia indirecta, etc. GRADE analiza de forma individualizada cada desenlace y cada factor pronóstico.

st0080 # ¿Qué es pronosticar?

p0605 Hacer un pronóstico consiste en conocer el futuro. Por lo menos, desde un punto de vista etimológico. A los efectos de este capítulo, los estudios pronósticos son aquellos que buscan averiguar qué le sucederá a un paciente afectado por una determinada circunstancia (una enfermedad, un factor de

ISBN: 978-84-9113-883-9; PII: B978-84-9113-883-9.00015-0; Autor: CABELLOSPAIN2021; Documento ID: 00015; Capítulo ID: c0075

riesgo, un tratamiento). Los estudios pronósticos tienen interés tanto para profesionales de la salud como para pacientes y sus familiares, políticos y encargados de tomar decisiones sobre salud (1).

p0610 El pronóstico forma parte de la tríada de la práctica clínica: diagnóstico-tratamiento-pronóstico. En la época hipocrática el pronóstico era el elemento más importante de los tres (se conocía poco de los elementos diagnósticos de las enfermedades, y el tratamiento en muchos casos se limitaba al conocido *primum non nocere*). En el siglo XX cobró importancia el diagnóstico (las opciones de tratamiento aún estaban bastante limitadas). Hoy día, los avances en el conocimiento de los mecanismos de las enfermedades permiten conocer más íntimamente los mecanismos de las enfermedades en un paciente concreto. Esto hace posible establecer un pronóstico más afinado para una determinada persona, con unas características particulares. Además, han surgido con fuerza técnicas para extraer y analizar cantidades ingentes de datos de salud *(big data)*.

p0615 Los términos referentes a los estudios pronósticos eran tradicionalmente confusos. Para solucionar este problema se creó la Estrategia en Investigación Pronóstica, en inglés PROGnosis RESearch Strategy, conocida por su acrónimo PROGRESS (2). En esta Estrategia se propone clasificar los estudios pronósticos en cuatro tipos distintos:

u0435 **PROGRESS tipo I:** estudios de pronóstico global. En estos estudios se analizan los resultados reales en muestras de pacientes con una determinada enfermedad o situación de salud de interés. Se llaman «globales» porque el resultado es una medida global, como los valores medios de una medida de la enfermedad. Por ejemplo, «puntuación en el MINIMENTAL test a los 12 meses» o «porcentaje que sigue sin poder trabajar a los 12 meses».

u0440 **PROGRESS tipo II:** estudios de factores pronósticos (3). Estudian qué características (o factores) se asocian con cambios en resultados globales para los individuos del estudio.

u0445 **PROGRESS tipo III:** modelos pronósticos (4). Evaluación de modelos matemáticos de predicción de riesgo que incorporan múltiples factores pronósticos.

u0450 **PROGRESS tipo IV:** predictores del efecto de un tratamiento (5). Estudian las características que predicen si un individuo responderá o no a un determinado tratamiento.

p0640 En este capítulo nos ceñiremos a las revisiones sistemáticas de estudios pronóstico tipo II (factores pronósticos). Un factor pronóstico puede definirse como una característica o variable del paciente, que está asociada a una determinada probabilidad de sufrir un resultado relevante (por ejemplo, sufrir un infarto). Para la estructura general de una revisión sistemática remitimos a los lectores al capítulo 12 («Lectura crítica de revisiones sistemáticas sobre estudios de prevención o tratamiento»). Muy resumidamente, los pasos de una revisión sistemática son (6):

u0455 ■ Protocolo.
u0460 ■ Búsqueda de los estudios individuales.
u0465 ■ Evaluación del riesgo de sesgo en cada estudio individual seleccionado.
u0470 ■ Extracción de los datos de cada estudio individual.
u0475 ■ Síntesis de la «evidencia» (si es posible).
u0480 ■ Diseminación de los resultados.

Conclusión

st0035

p0650 La lectura crítica de revisiones sistemáticas de estudios pronósticos posee unas peculiaridades respecto a otro tipo de revisiones sistemáticas. Se deberá tener en cuenta el tipo de estudio de pronóstico analizado. La pregunta clínica tiene un formato PICO-TA. La herramienta CHARMS (CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) es una lista con los ítems que deben extraerse de los estudios individuales que se incluyen en una revisión sistemática de estudios pronósticos. La herramienta QUIPS (QUality In Prognosis Studies) se emplea para evaluar el riesgo de sesgo en estas revisiones. GRADE es una herramienta que mide de modo independiente la certidumbre en la evidencia y la fuerza de la evidencia. Las revisiones de estudios pronósticos tipo II enlazan con otros estudios tipo III y tipo IV (modelos

ISBN: 978-84-9113-883-9; PII: B978-84-9113-883-9.00015-0; Autor: CABELLOSPAIN2021; Documento ID: 00015; Capítulo ID: c0075

pronósticos y estudios predictores a tratamiento, y con estudios de medicina personalizada, que influirán de forma determinante en el desenlace vital de nuestros pacientes.

st0090 ## Artículo

p0680 Fernando SM, Tran A, Wei Cheng, et al. Pre-arrest and intra-arrest prognostic factors associated with survival after in-hospital cardiac arrest: systematic review and meta-analysis. BMJ. 2019;367:l6373. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6891802/.

st0095 ## Plantilla CASPe contestada para este artículo concreto

p0685 En el cuadro 15.1 se muestra la plantilla CASPe contestada para este artículo concreto.

b0010 **CUADRO 15.1 ■ Evaluación crítica del artículo propuesto (plantillas CASPe)**

p0165 **A) ¿Los resultados de la revisión son válidos?**

Preguntas «de eliminación»			
1. ¿Se hizo la revisión sobre un tema claramente definido? *PISTA: un tema debe ser definido en términos de (PICO-TA):* • *La **P**oblación de estudio.* • *Los **Í**ndices pronósticos.* • ***C**omparadores (si procede).* • *Los desenlaces considerados (**O**utcomes).* • ***T**iempo (factor y desenlace).* • ***Á**mbito.*	Sí ✓	No sé	No
	No se describe específicamente la pregunta en formato PICO-TA (en inglés, PICO-TS). No obstante, del texto se recoge: • *Pacientes:* pacientes ingresados que han sufrido un paro cardíaco. Se incluyen los estudios con al menos un 80% de «adultos», considerados como al menos de 16 años. • *Índice pronóstico:* no se menciona ninguno específico. Factores preparada y factores intraparada. • *Comparador/control:* recuérdese que el comparador hace referencia a otro factor pronóstico con el que se compara el factor a estudio; mientras que el control es otro factor que se utiliza en el ajuste de la predicción con el factor de estudio. En esta revisión no existe un *comparador*, es decir, se evalúa el efecto pronóstico de múltiples factores sin un objetivo comparativo entre ellos. En cuanto a factores de *control* (o ajuste), en el manuscrito no se hace referencia a ellos. Sin embargo, en el protocolo se ha predefinido que los estudios de factores pronósticos incluidos en la revisión deben haber ajustado las estimaciones del efecto al menos por edad y sexo. • Outcome: mortalidad intrahospitalaria a los 28 o 30 días. • Timing: se evalúan por separado los factores medidos previos al paro cardíaco y los factores medidos durante el paro cardíaco. El tiempo de desenlace se especifica en el desenlace *(outcome)* de interés. • *Ámbito* (Setting): ámbito hospitalario. CONCLUSIÓN: SÍ.		

ISBN: 978-84-9113-883-9; PII: B978-84-9113-883-9.00015-0; Autor: CABELLOSPAIN2021; Documento ID: 00015; Capítulo ID: c0075

2. ¿Buscaron los autores el tipo de artículos adecuado? *PISTA: el mejor «tipo de estudio» es el que:* u0225 • *Se dirige a la pregunta objeto de la revisión y tiene el diseño apropiado.* u0230 • *Tipo de estudio de pronóstico.*	Sí ✓	No sé	No
	Los autores hacen referencia a la inclusión de estudios observacionales con diseño retrospectivo y prospectivo, ensayos clínicos aleatorizados y ensayos cuasi aleatorizados. Los estudios con un diseño prospectivo son considerados los mejores para responder a una pregunta de factores pronóstico. Estos permiten una mejor definición de los criterios de inclusión. La recogida de información, basal y durante el seguimiento, es más completa y estandarizada, tanto en forma como en la definición de factores y desenlaces bajo estudio. Esa diversidad contribuye a la heterogeneidad. CONCLUSIÓN: SÍ.		
Preguntas detalladas			
3. ¿Crees que estaban incluidos los estudios importantes y pertinentes? *PISTAS DE LA BÚSQUEDA:* u0235 • *¿Qué bases de datos bibliográficas se han usado? ¿Qué estrategia de búsqueda?* u0240 • *Seguimiento de las referencias.* u0245 • *Contacto personal con autores.* u0250 • *Búsqueda de estudios no publicados.* u0255 • *Idiomas distintos del inglés.* *PISTAS DE LA SELECCIÓN:* u0260 • *Criterios de inclusión/exclusión.* u0265 • *Selección estudios. 1/2.* u0270 • *Exacción de datos ½ (usaron CHARMS-PF).*	Sí ✓	No sé	No
	Se llevó a cabo la búsqueda en Medline, PubMed, Embase, Scopus, Web of Science, y la Cochrane Database of Systematic Reviews. Se utilizó la herramienta *Related Articles* («artículos relacionados») de PubMed para ampliar la búsqueda. La estrategia de búsqueda, que se presenta en el material suplementario, fue diseñada por una bibliotecaria experimentada en ciencias de la salud. Esta combina términos específicos de la población bajo revisión, tales como *cardiac arrest,* y relacionados con la investigación en pronóstico, como *prognostic* o *risk.* Los autores de la revisión contactaron con los autores de correspondencia de los estudios primarios, cuando estos no reportaron las estimaciones de los *odds ratios,* ajustados o crudos, o los datos necesarios a partir de los cuales poder calcularlos. Aunque en la estrategia de búsqueda no hay limitaciones idiomáticas, solo se incluyen artículos en inglés. Aunque no consideramos que se trate de un porcentaje relevante, sería interesante conocer cuántos estudios fueron descartados por estar escritos en otro idioma, en nuestra opinión presumiblemente pocos. CONCLUSIÓN: SÍ.		
4. ¿Crees que los autores de la revisión han hecho suficiente esfuerzo para valorar la calidad de los estudios incluidos? *PISTA 1: QUIPS «riesgo de sesgo» depende de 6 dominios: población, atrición, medición de factores pronósticos, medición de desenlaces, confusión y análisis estadístico correcto.*	Sí ✓	No sé	No
	Los autores han valorado la calidad de los estudios mediante la herramienta QUIPS (QUality In Prognosis Studies), la herramienta recomendada para evaluar riesgo de sesgo en este tipo de revisión sistemática. El riesgo de sesgo se evalúa usando 31 preguntas divididas en seis dominios (población, atrición, medición de los factores pronósticos, medición del desenlace, factores de ajuste y análisis estadístico), y para dominio los estudios se clasifican en bajo, moderado y alto riesgo de sesgo. Se presentan los resultados en la tabla 7 del material adicional. Véase como, en la mayoría de los dominios, el riesgo de sesgo es bajo. En otros, como el sesgo de atrición en Ballew, es moderado. CONCLUSIÓN: SÍ.		

ISBN: 978-84-9113-883-9; PII: B978-84-9113-883-9.00015-0; Autor: CABELLOSPAIN2021; Documento ID: 00015; Capítulo ID: c0075

5. Si los resultados de los diferentes estudios han sido mezclados para obtener un resultado «combinado», ¿era razonable hacer eso? *PISTA: heterogeneidad (I^2) puede ser:* u0275 • *Clínica.* u0280 • *Metodológica.* u0285 • *Estadística .*	Sí ✓	No sé No
	La heterogeneidad fue evaluada usando el estadístico I^2, el test χ^2 para la homogeneidad y, visualmente, mediante los *forest plots.* El estadístico I^2 era muy alto para varios de los metaanálisis realizados. Por ejemplo, en la evaluación de los factores pronóstico preparada: sexo masculino I^2 = 66% o edad > 70 I^2 = 69%; y de los factores intraparada: diagnóstico de síndrome coronario agudo I^2 = 99%. Los autores indican que la heterogeneidad es debida a que en el metaanálisis se combinan estudios de grandes registros nacionales con una variabilidad pequeña de la estimación del efecto y estudios pequeños con mayor variabilidad, y justifican su uso basado en el solapamiento de las estimaciones puntuales y los correspondientes intervalos de confianza al 95%. Sin embargo, y ante la evidente presencia de heterogeneidad entre estudios, hubiera sido de interés para el lector que en el metaanálisis se hubieran presentado los intervalos de predicción, que informan del potencial valor verdadero del efecto en un nuevo estudio. Un análisis de sensibilidad, excluyendo del metaanálisis los estudios pequeños de gran variabilidad, también podría resultar interesante para el lector, de modo que permita examinar esas potenciales causas de heterogeneidad.	
B) ¿Cuáles son los resultados?		
6. ¿Cuál es el resultado global de la revisión? *PISTA: considera:* u0290 • *Valora para los desenlaces positivos y también los negativos.* u0295 • *¿Cuáles son los resultados para cada desenlace?* u0300 • *¿Cómo están expresados los resultados RR, HR, etc.)?* u0305 • *¿Muestran gráficos* forest plots*?*	Los resultados para el análisis principal que evalúa la asociación entre los factores pronósticos preparada e intraparada con la *odds* de supervivencia, se presentan en la tabla 2 y las figuras 2-3 (*forest plots* para los factores pronóstico preparada) y 4-5 (*forest plots* para los factores pronóstico intraparada). Para cada factor pronóstico estudiado se indican los estudios incluidos en el metaanálisis, el tamaño del efecto en términos de *odds ratios* y sus correspondientes intervalos de confianza al 95%, la heterogeneidad mediante el estadístico I^2 y el grado de certeza mediante la evaluación GRADE (tabla 15.2). **Factores preparada:** El sexo masculino (OR = 0,84 [0,73 a 0,95] con grado de certeza moderado), la edad avanzada (OR = 0,50 [0,40 a 0,62] para edad mayor a 60 años y OR = 0,42 [0,18 a 0,99] para edad mayor a 70 años, ambos con grado de certeza bajo), la existencia de una neoplasia concomitante (OR = 0,57 [0,45 a 0,71] con grado de certeza alto) y la existencia de enfermedad renal crónica (OR = 0,56 [0,40 a 0,78] con grado de certeza alto) se relacionaron con un peor pronóstico. Otras comorbilidades (insuficiencia cardíaca congestiva, enfermedad pulmonar obstructiva crónica y diabetes mellitus) y el diagnóstico de sepsis a la admisión se asociaron a un peor pronóstico en estudios individuales (sin metaanálisis). **Factores intraparada:** Los paros cardíacos ante testigos (OR = 2,71 [2,17 a 3,38]), la monitorización (OR = 2,23 [1,41 a 3,52]), los paros cardíacos diurnos (con plantilla al completo) (OR = 1,41 [1,20 a 1,66]), la fibrilación ventricular (OR = 3,68 [2,68 a 5,05]) y la taquicardia ventricular (OR = 3,76 [2,95 a 4,78]) se relacionaron con un mejor pronóstico. Todos ellos con alto grado de certeza. Mientras que la asistolia (OR = 0,42 [0,32 a 0,56] con grado de certeza alto), la intubación traqueal (OR = 0,54 [0,42 a 0,70] con grado de certeza moderado) y la duración prolongada de las maniobras de resucitación cardiopulmonar (OR = 0,12 [0,07 a 0,19] con grado de certeza alto) se relacionaron con un peor pronóstico.	

ISBN: 978-84-9113-883-9; PII: B978-84-9113-883-9.00015-0; Autor: CABELLOSPAIN2021; Documento ID: 00015; Capítulo ID: c0075

7. Para el conjunto de los estudios (en cada desenlace concreto) • *¿Cuál es la precisión de los resultados?* • *¿Son consistentes los resultados de los estudios para cada desenlace?* • *¿Es indirecta la evidencia en algún desenlace?*	En la revisión sistemática de ejemplo, dado que son varios los factores pronósticos bajo estudio, las siguientes cuestiones deben discutirse para cada factor concreto. • **¿Cuál es la precisión de los resultados?**Para evaluar de forma crítica la precisión de los resultados debemos fijarnos en los intervalos de confianza de la estimación puntual del efecto. Por ejemplo, para el factor pronóstico *historia de malignidad* (preparada) la estimación puntual del *pooled odds ratio* es 0,57, con un intervalo de confianza al 95% relativamente ajustado (preciso), entre 0,45 y 0,71. Sin embargo, para el factor *edad* ≥ *70* (preparada), la estimación puntual del *pooled odds ratio* es 0,42, y su intervalo de confianza mucho más holgado (impreciso), entre 0,18 y 0,99. Ver tabla 2. • **¿Son consistentes los resultados de los estudios para cada desenlace?**La consistencia de los resultados depende de la heterogeneidad entre los estudios. Esta se puede valorar a partir de los valores del estadístico I^2, los test estadísticos de heterogeneidad u observando el grado de solapamiento entre los intervalos de confianza de los estudios identificados. En el caso de disponer de ellos, también podríamos ayudarnos de los intervalos de predicción. El valor del estadístico I^2 en todos los factores pronóstico metaanalizados excede del 50%, indicando un importante grado de heterogeneidad entre los estudios; además, las potenciales causas de heterogeneidad no han sido exploradas mediante análisis de subgrupos. Por tanto, la consistencia de los resultados es moderada. • **¿Es indirecta la evidencia en algún desenlace?** Cuando la definición del desenlace, el factor pronóstico o la población bajo revisión diverge entre los estudios incluidos, podría ser un síntoma de evidencia indirecta.
C) ¿Son los resultados aplicables en tu medio?	
8. ¿Se pueden aplicar los resultados en tu medio? *PISTA: considera si:* • *Los pacientes cubiertos por la revisión pueden ser suficientemente diferentes de los de tu área.* • *Tu medio es muy diferente a los del estudio.*	Sí ✓ No sé No Los pacientes y el ámbito donde se ha elaborado la revisión sistemática son, en principio, similares a los que encontraríamos en otros hospitales de nuestro entorno. CONCLUSIÓN: parece que sí, aunque desconocemos elementos como infraestructura, protocolos o entrenamiento del personal de los centros.
9. ¿Se han considerado todos los resultados necesarios para tomar una decisión? • *¿Qué te gustaría saber además de esto?*	Sí No sé ✓ No Sería interesante comprobar el papel de la telemedicina para detección de paradas cardíacas, que parece que desempeñará un papel de importancia creciente. También convendría saber cuál es el nivel de entrenamiento y formación del personal hospitalario en esta cuestión. Otra variable que se debe tener en cuenta es qué tipo de planta hospitalaria y sus características (medica/quirúrgica, etc.).

ISBN: 978-84-9113-883-9; PII: B978-84-9113-883-9.00015-0; Autor: CABELLOSPAIN2021; Documento ID: 00015; Capítulo ID: c0075

10. ¿Crees que hay alguna medida que tomar en tu caso?

u0355 • *Aunque no esté planteado explícitamente en la revisión, ¿qué opinas?*

Sí ✓ No

El manejo de las paradas cardíacas es complejo y depende de varios factores, que incluyen las características estructurales del propio hospital, las características y el entrenamiento del personal y los propios protocolos de manejo de las paradas.

Conocer los factores pronósticos relacionados con la *odds* de supervivencia es útil para afinar el diseño de protocolos y guías de actuación. Además, a partir de los resultados de la revisión tenemos un interesante punto de partida para el desarrollo de un modelo pronóstico que nos permita estimar de manera individualizada la probabilidad de supervivencia a una parada cardíaca según las características del paciente, así como la probabilidad de parada en un paciente cuando ingresa *(pre-arrest)*, con la finalidad de prevenir o jerarquizar el riesgo.

t0015 TABLA 15.2 ■ **Tabla sumario de resultados GRADE por factores**

Factor estudiado y n.º de estudios	Riesgo de sesgo	Imprecisión	Inconsistencia*	Evidencia indirecta	Certeza en la evidencia**	Estimador IC 95%
«Parada» presenciada por terceros (4 estudios)	No importante	No importante	No importante I^2 = 68%	No importante	Certeza GRADE alta	2,71 (2,17-3,38)
Paciente monitorizado (6 estudios)	No importante	No importante	No importante I^2 = 97%	No importante	Certeza GRADE alta	2,23 (1,41-3,52)
Parada en horario diurno (5 estudios)	No importante	No importante	No importante I^2 = 94%	No importante	Certeza GRADE alta	1,41 (1,20-1,66)
Ritmo inicial susceptible de desfibrilación (12 estudios)	No importante	No importante	No importante I^2 = 96%	No importante	Certeza GRADE alta	5,28 (3,78-7,39)
Intubación durante la parada (5 estudios)	No importante	No importante	No importante I^2 = 73%	Importante***	Certeza GRADE moderada	0,54 (0,42-0,70)
Duración resucitación > 15 min (2 estudios)	Importante	No importante	No importante I^2 = 27%	No importante	Certeza GRADE alta	0,12 (0,07-0,19)

Explicaciones: *A pesar de altos valores de I^2 hay alto grado de solapamiento entre las estimaciones puntuales y los intervalos de confianza. **La mayor parte del peso en la estimación del efecto agrupado proviene de estudios de bajo riesgo de sesgo, excepto para el factor «duración de la resucitación», que proviene de estudios de moderado riesgo de sesgo. ***Tiempos de intubación variables y no están claras otras variables de confusión que contribuyen a si el paciente está o no intubado.

La tabla (tomada del artículo) presenta algunas discrepancias respecto a la presentada por los autores de la revisión en el material adicional (tabla suplementaria 9). Esto sugiere que se ha llevado a cabo un análisis de sensibilidad o que por alguna otra razón se ha descartado algún estudio en el análisis final presentado en el manuscrito. En cualquier caso, en ambas tablas la certeza en la evidencia de las estimaciones no cambia (es alta), y aunque los OR difieren ligeramente (como es lógico con distinto número de estudios incluidos en el análisis), la dirección del efecto es consistente en ambas tablas. Por lo tanto, no existen discrepancias esenciales.

No obstante, se ha escrito a los autores y editores para la justificación o corrección de estas inconsistencias.

ISBN: 978-84-9113-883-9; PII: B978-84-9113-883-9.00015-0; Autor: CABELLOSPAIN2021; Documento ID: 00015; Capítulo ID: c0075

Bibliografía

bib0010 1. Riley RD, van der Windt DA, Croft P, Moons KGM. Prognosis Research in Health Care: Concepts, Methods, and Impact. Oxford: Oxford University Press; 2019.

bib0015 2. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. PROGRESS Group. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ 2013;346:e5595.

bib0020 3. Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al. PROGRESS Group. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Med 2013;10(2):e1001380.

bib0025 4. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10(2):e1001381.

bib0030 5. Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, et al. PROGRESS Group. Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ 2013;346:e5793.

bib0035 6. Riley RD, Moons KGM, Snell KIE, Ensor J, Hooft L, Altman DG, et al. A guide to systematic review and meta-analysis of prognostic factor studies. BMJ 2019;364:k4597.

bib0040 7. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PloS One 2012;7(2):e32844.

bib0045 8. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014;11(10):e1001744.

bib0050 9. Hayden JA, van der Windt DA, Cartwright JL, et al. Assessing bias in studies of prognostic factors. Ann Intern Med 2013;158(4):280-6.

bib0055 10. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7(3):177-88.

bib0060 11. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. Stat Med 2001;20(24):3875-89.

bib0065 12. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. BMC 2015;15:99.

bib0070 13. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924-6.

Cómo citar este capítulo:

bib0075 Maldonado M, Fernández Félix BM, Cabello JB. Lectura crítica de revisiones sistemáticas de estudios de pronóstico o riesgo. En: Cabello Juan B, editor. Lectura crítica de la evidencia clínica, 2.ª ed. Barcelona: Elsevier; 2022. p. X-X.

ISBN: 978-84-9113-883-9; PII: B978-84-9113-883-9.00015-0; Autor: CABELLOSPAIN2021; Documento ID: 00015; Capítulo ID: c0075