UNIVERSIDAD MIGUEL HERNÁNDEZ DE ELCHE

FACULTAD DE MEDICINA

Programa de Doctorado en Salud Pública, Ciencias Médicas y Quirúrgicas

UN MÉTODO PARA VALIDAR SISTEMAS DE PUNTOS BASADOS EN MODELOS DE REGRESIÓN LOGÍSTICA PARA PREDECIR EVENTOS BINARIOS CON LA METODOLOGÍA BOOTSTRAP Y A TRAVÉS DE UNA APP PARA SMARTPHONE

TESIS DOCTORAL REALIZADA POR:

D. David Manuel Folgado De la Rosa

DIRIGIDA POR LOS PROFESORES DOCTORES:

D. Vicente Francisco Gil Guillén (Director)

D. Antonio Palazón Bru (Codirector)

San Juan de Alicante, 2021

Esta tesis doctoral está basada en dos artículos científicos y se presenta por compendio de publicaciones:

Palazón-Bru A, Folgado-de la Rosa DM, Cortés-Castell E, López-Cascales MT, Gil-Guillén VF. Sample size calculation to externally validate scoring systems based on logistic regression models. *PLoS One* 2017; **12**: e0176726. doi: 10.1371/journal.pone.0176726.

Folgado-de la Rosa DM, Palazón-Bru A, Gil-Guillén VF. A method to validate scoring systems based on logistic regression models to predict binary outcomes via a mobile application for Android with an example of a real case. *Comput Methods Programs Biomed* 2020; **196**: 105570. doi: 10.1016/j.cmpb.2020.105570.

Los Profesores Doctores D. Vicente Francisco Gil Guillén y D. Antonio Palazón Bru, como Director y Codirector de Tesis Doctoral, respectivamente

CERTIFICAN:

Que el trabajo "UN MÉTODO PARA VALIDAR SISTEMAS DE PUNTOS BASADOS EN MODELOS DE REGRESIÓN LOGÍSTICA PARA PREDECIR EVENTOS BINARIOS CON LA METODOLOGÍA BOOTSTRAP Y A TRAVÉS DE UNA APP PARA SMARTPHONE", realizado por D. David Folgado De la Rosa ha sido llevado a cabo bajo nuestra dirección/codirección y se encuentra en condiciones de ser leído y defendido como Tesis Doctoral.

Lo que firmamos para los oportunos efectos en Sant Joan d' Alacant a Treinta de Enero de Dos Mil Veintiuno.

Prof. Dr. D. Vicente
Francisco Gil Guillén
Director

Prof. Dr. D. Antonio
Palazón Bru
Codirector

La Comisión Académica del Programa de Doctorado en Salud Pública, Ciencias Médicas y Quirúrgicas de la Universidad Miguel Hernández representada por su Coordinador, el Prof. Dr. D. Vicente Francisco Gil Guillén,

AUTORIZA:

La presentación y defensa como Tesis Doctoral del trabajo "UN MÉTODO PARA VALIDAR SISTEMAS DE PUNTOS BASADOS EN MODELOS DE REGRESIÓN LOGÍSTICA PARA PREDECIR EVENTOS BINARIOS CON LA METODOLOGÍA BOOTSTRAP Y A TRAVÉS DE UNA APP PARA SMARTPHONE" realizado por D. David Folgado De la Rosa bajo la dirección/codirección de los Profesores Doctores D. Vicente Francisco Gil Guillén y D. Antonio Palazón Bru, respectivamente.

Lo que firmamos para los oportunos efectos en Sant Joan d' Alacant a Treinta de Enero de Dos Mil Veintiuno.

Prof. Dr. D. Vicente Francisco Gil Guillén

Coordinador del Programa de Doctorado Salud Pública, Ciencias Médicas y Quirúrgicas de la Universidad Miguel Hernández

Dedicatoria

A mis padres, a mi familia por completo, amigos y todos los que alguna vez han creído en mí. También a los que no lo han hecho.

Agradecimientos

A D. Antonio Palazón Bru por tu tiempo, tus conocimientos, tus ánimos incondicionales, los ratos vividos y por orientar y guiarme por este viaje durante tantos años. Este trabajo, nos deja unidos de por vida con una afianzada amistad, por lo que, como dice el refrán, lo que haya unido El Romeral, los burpees y la ciencia, que no lo separe nadie.

A D. Vicente Francisco Gil Guillén, por tus incansables muestras de ánimo y tu sentido del humor durante esta dirección de tesis.

A Dña Ana María Martínez Díaz por esas risas con deje gaditano entre esos estreses, seminarios y exposiciones.

Oh! inmortal Poseidón el del furioso tridente.

A ti me encomiendo en esta difícil empresa .

Propicia que este velero llegue a buen puerto.

Permíteme llevar a cabo los designios de Afrodita,

nacida de las olas.

Oh! Atenea, augusta entre las diosas,

Haz florecer el jardín, trae la ciencia…

Texto adaptado de **Kase.O**

ÍNDICE

1. Resumen en inglés

Predicting a binary event in a particular population with a points-based system based on a logistic regression model requires its prior validation. This validation necessitates, first, an adequate sample size and, second, application of the mathematical techniques recommended by international consensus.

Concerning the first requirement, the recommendations are that the sample size have at least 100 events and 100 non-events, independently of the model being validated. However, scientific studies have shown that certain factors can influence the sample size, so having a fixed value for this purpose does not seem to be sensible. Consequently, this doctoral thesis has developed an algorithm to calculate the sample size in points-based scoring systems obtained through logistic regression models. This algorithm is based on methods widely used in the scientific literature, such as the ROC curve or bootstrapping.

Regarding the second requirement, the most correct way is to calculate the discrimination and calibration with bootstrapping. The discrimination can be done using the area under the ROC curve and the calibration with a smoothed calibration plot (most recommended method). As this is not an easy task, a method is suggested to construct an application for an Android phone that does it.

Both methods have been used on a simulated data set from a model to predict mortality in Intensive Care Units. This enables each step to be understood so that it can be applied with other points-based scoring systems to predict binary events.

2. Resumen en castellano

Para emplear un sistema de puntos basado en un modelo de regresión logística para predecir un evento binario en una determinada población, es necesaria su validación. Para ello, debemos tener un tamaño muestral suficiente y aplicar las técnicas matemáticas recomendadas por los consensos internacionales.

En lo referente al primer punto, se ha recomendado tener un tamaño muestral con al menos 100 eventos y 100 no eventos, independientemente del modelo que estemos validando. Sin embargo, estudios científicos han demostrado que ciertos factores influyen sobre este tamaño, por lo que parece que no tiene mucho sentido tener un valor fijo para este propósito. En consecuencia, en esta tesis doctoral se ha desarrollado un algoritmo para calcular el tamaño muestral en sistemas de puntos obtenidos a través de modelos de regresión logística. Este algoritmo se basa en métodos utilizados ampliamente en la literatura científica, como son la curva ROC o el bootstrapping.

Con respecto al segundo y último punto, la forma más correcta de ello es calcular a través de bootstrapping la discrimación y la calibración. La discriminación se puede abordar a través del área bajo la curva ROC y la calibración mediante la representación del gráfico de calibración suavizado (método más recomendado). Esto no es una tarea sencilla, por lo que se plantea la elaboración de una metodología para construir una aplicación para teléfono móvil en Android que la realice.

Ambos métodos se han aplicado sobre un conjunto de datos simulados perteneciente a un modelo para predecir mortalidad en las Unidades de Cuidados Intensivos. De esta forma se puede comprender cada una de las etapas, con el fin de poder aplicarse para otros sistemas de puntos para predecir eventos binarios.

3. Introducción

3.1 El modelo de regresión logística

En biomedicina, y por extensión en la vida cotidiana, nos encontramos con sucesos que pueden tomar sólo dos valores excluyentes entre sí (binarios). Por ejemplo, la mortalidad de un paciente, la recurrencia de una enfermedad o si lloverá mañana o no. Si se requiere conocer la probabilidad de que ocurra el evento, podemos tratar de estimarla si conocemos factores que influyen en la misma.[1] Un caso práctico actual sería un paciente con infección por coronavirus de 2019, ya que tener mayor edad, hipertensión arterial o ser varón, influyen de forma negativa en el pronóstico.[2] Una de las herramientas más utilizadas para analizar estas relaciones, es el modelo de regresión logística binario.[1]

3.1.1 Univariante

Siguiendo con el ejemplo del coronavirus de Wuhan, supongamos que queremos explicar a través de un modelo de estas características, el pronóstico de un paciente conociendo su edad, es decir, mediante una única variable explicativa (caso univariante). Vamos a formular este caso de manera generalizada: en una población determinada, sea X, esa variable independiente y Z la variable binaria dependiente, la cual sigue una distribución Bernoulli, pues tan solo puede tomar dos valores, ocurriendo uno de ellos como probabilidad π y el otro $1-\pi$. La esperanza de este tipo de distribución es π, la cual estamos interesados en conocer. Esto lo haremos a través de valores conocidos de la variable X (x): $E(Z|x) = \pi(x)$.

El modelo de regresión logística logra este cometido a través de la siguiente expresión:

$$g(\pi(x)) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 \cdot x,$$

donde g es la función logit, y β_0 y β_1 corresponden a los coeficientes del modelo. β_0 corresponde con la constante o intercept, mientras que β_1 es el coeficiente asociado a x. De esta expresión se puede deducir la siguiente, en donde conociendo los valores de x, podemos obtener $\pi(x)$:[1]

$$E(Z|x) = \pi(x) = \frac{e^{(\beta_0 + \beta_1 \cdot x)}}{1 + e^{(\beta_0 + \beta_1 \cdot x)}}$$

3.1.1.2 Estimación

Los coeficientes del modelo, β_0 y β_1, son parámetros poblacionales, los cuales vamos a aproximar o estimar a través de una muestra de sujetos obtenidos de la población. Para ello, recordemos que la variable dependiente Z sigue una función Bernoulli, la cual tiene una función de densidad $f(z_i) = \pi_i^{z_i}(1-\pi_i)^{1-z_i}$, donde para un sujeto i, z_i nos indica la presencia o ausencia del suceso binario que estamos valorando ($1 = Sí,\ 0 = No$) y π_i es la probabilidad de que ese evento ocurra. En consecuencia, la función de densidad vale π_i cuando sucede el evento y $1-\pi_i$ cuando no.

Supongamos que tenemos n sujetos ($i = 1, 2, ..., n$), se define la función de verosimilitud como el producto de todas sus funciones de densidad, es decir:

$$\prod_{i=1}^{n} f(z_i) = \prod_{i=1}^{n} \pi_i^{z_i} (1-\pi_i)^{1-z_i}.$$

Si tomamos el logaritmo de esta expresión, obtenemos la función de log-verosimilitud:

$$log(\prod_{i=1}^{n} \pi_i^{z_i} (1-\pi_i)^{1-z_i}) = \sum_{i=1}^{n} \left(z_i\, log(\pi_i) + (1-z_i) log(1-\pi_i) \right).$$

Ahora, vamos a aplicar el modelo de regresión logística, es decir, sustituir π_i por $\pi_i(x_i)$

$$\sum_{i=1}^{n} \left(z_i\, log(\pi_i(x_i)) + (1-z_i) log(1-\pi_i(x_i)) \right) =$$
$$\sum_{i=1}^{n} \left(z_i\, log(\frac{e^{(\beta_0 + \beta_1 x_i)}}{1+e^{(\beta_0 + \beta_1 x_i)}}) + (1-z_i) log(1-\frac{e^{(\beta_0 + \beta_1 \cdot x_i)}}{1+e^{(\beta_0 + \beta_1 \cdot x_i)}}) \right).$$

Si nosotros maximizamos esta expresión en función de β_0 y β_1, pues el resto de elementos es conocido (z_i y x_i), obtendremos la estimación de los mismos ($\widehat{\beta}_0$ y $\widehat{\beta}_1$). Este procedimiento no tiene una expresión cerrada para su cálculo, por lo que se utilizan métodos numéricos como el de Newton-Raphson.[1]

3.1.1.2 Significancia

Ahora, una vez sabemos estimar los coeficientes del modelo, nos surge la pregunta de si éstos son significativos en nuestra población, es decir, si tienen influencia sobre nuestra variable dependiente o lo que es equivalente, que tengan un valor no nulo. En caso de β_1, esto equivale a que el predictor asociado, X, tiene una influencia sobre la variable dependiente, Z. Para este propósito, aunque existen otros tipos de tests, como el test de Wald, se recomienda utilizar el test de máxima verosimilitud.[3,4]

Para la realización de esta prueba vamos a calcular la función de verosimilitud con los valores estimados $\widehat{\beta}_0$ y $\widehat{\beta}_1$, y por otro lado vamos a estimar un nuevo modelo sin el coeficiente que queremos ver su significancia, en este caso β_1. A continuación calculamos la función de verosimilitud en ambos modelos y construimos la expresión:

$$G = -2 \log\left(\frac{Verosimilitud\ sin\ \beta_1}{Verosimilitud\ con\ \beta_1} \right).$$

Ésta, suponiendo que β_1 tenga un valor nulo en la población, se distribuye como una χ^2 con un grado de libertad, por lo que con este test podemos determinar la significancia de β_1, es decir, estamos contrastando $\beta_1 \neq 0$.[1]

3.1.2 Multivariante

Vamos a tratar de extender el ejemplo del coronavirus de Wuhan,[2] ya que es posible que el pronóstico de un paciente se vea afectado por más de un factor al mismo tiempo (edad, hipertensión arterial, sexo…). En contraposición al caso univariante visto anteriormente, donde únicamente valorábamos la edad del paciente, esto se conoce como modelo de regresión logística multivariante.

Por la tanto, sea Z nuestra variable binaria dependiente y el conjunto $\{X_1, ..., X_p\}$ las variables independientes. El modelo de regresión logística, conocidos valores de las variables independientes, tendría la siguiente expresión:

$$g(\pi(x)) = ln\left[\frac{\pi(x)}{1-\pi(x)} \right] = \beta_0 + \beta_1 \cdot x_1 + ... + \beta_p \cdot x_p.$$

Por extensión, podemos seguir los pasos anteriores para obtener la estimación y la significancia, obteniendo fórmulas similares.[1]

3.1.3 Validación

Independientemente del método estadístico utilizado para desarrollar el modelo, debemos valorar la validación, es decir si éste predice adecuadamente nuestra variable dependiente en un conjunto de sujetos. Si este conjunto es aquél donde se ha desarrollado el modelo, este proceso se conoce como validación interna, en caso contrario, validación externa. En ambas, se valora tanto la discriminación como la calibración, las cuales desarrollaremos a continuación.[5-9]

3.1.3.1 Discriminación

Volviendo al ejemplo del coronavirus, podemos intuir que un paciente con mayor edad es más propenso al fallecimiento que uno más joven, es decir, la edad nos está diferenciando o discriminando qué sujeto tiene más riesgo de fallecer. En términos generales, se refiere a cómo de bien el modelo diferencia entre aquéllos con y sin el evento. Ésta se evalúa normalmente mediante el estadístico c, el cual es equivalente al área bajo la curva ROC (acrónimo de Receiver Operating Characteristic).[5,10]

3.1.3.1.1 Curva ROC

Partiendo con la notación anterior, definimos $E = \{i : z_i = 1,\ i = 1,\ ...,\ n\}$ y $\overline{E} = \{i : z_i = 1,\ i = 1,\ ...,\ n\}$, de forma que $E \cup \overline{E} = \{1, ..., n\}$ y $E \cap \overline{E} = \varnothing$. La curva ROC se define como la unión de los siguientes puntos en un gráfico cartesiano:

$$\left(1 - \frac{|i \in \overline{E} : x_i < x|}{|\overline{E}|},\ \frac{|i \in E : x_i \geq x|}{|E|}\right) x \in R(x),$$

siendo $|\cdot|$ la función cardinal de un conjunto dado, es decir, el número de elementos de dicho conjunto.

Ahora, vamos a calcular el área bajo la curva ROC (ABC) en el espacio [0,1] *x* [0,1]. Sea $j \in E\ y\ \overline{E}$, se define:

$$(,\ ) = \left\{1,\ si\ x_j > x_l; 1/2,\ si\ x_j = x_l; 0,\ si\ x_j < x_l\right\}.$$

El cálculo del ABC se obtiene con la siguiente expresión:

$$ABC = \frac{1}{|E| \cdot |\overline{E}|} \cdot \sum_{j \in E} \sum_{l \in \overline{E}} S(j, l),$$

siendo la fórmula cerrada para su error estándar:

$$EE(ABC) = \sqrt{\frac{ABC \cdot (1 - ABC) + (|E| - 1) \cdot (Q_1 - ABC^2) + (|\overline{E}| - 1) \cdot (Q_2 - ABC^2)}{|E| \cdot |\overline{E}|}},$$

donde $Q_1 = \frac{ABC}{(2 - ABC)}$ y $Q_2 = \frac{2 \cdot ABC^2}{(1 + ABC)}$.

A la hora de interpretar el valor del ABC, el cual es un número entre 0 y 1, cuanto más cercano a la unidad sea, nos estará indicando que la variable X diferencia en mayor medida entre qué sujeto tiene un evento y qué sujeto no lo tiene.

En el caso del modelo de regresión logística univariante, X será nuestra variable independiente que definimos en las secciones anteriores, mientras que en el caso multivariante, se corresponde con las probabilidades pronosticadas del modelo, obtenidas conociendo el valor de los predictores y de la estimación de los coeficientes.[11]

3.1.3.2 Calibración

Diremos que un modelo predictivo está bien calibrado si los resultados obtenidos por éste se asemejan a la realidad. Existen varios métodos para determinar la calibración, siendo de los más comunes la representación gráfica cartesiana de las probabilidades observadas frente a las pronosticadas.[5]

Dichos gráficos son, normalmente, complementados por un test estadístico, como el test de Hosmer-Lemeshow. Sin embargo, este tipo de tests han sido frecuentemente criticados, debido a la limitada potencia estadística cuando hay poca calibración y a ser muy sensibles en muestras grandes. Además, el test de Hosmer-Lemeshow no proporciona información sobre el rango de valores en los que existe descalibración, en caso de que rechacemos la hipótesis nula de este test.[5,7,8,12,13]

3.1.3.2.1 El test de Hosmer-Lemeshow

El test de Hosmer–Lemeshow es un test estadístico para determinar lo bien que un modelo de regresión logística se ajusta y es muy utilizado en modelos de predicción de riesgo. El test valora si las probabilidades del evento observado y esperado son parecidas en diferentes subgrupos de la muestra sobre la que se está validando el modelo.[1]

3.1.3.2.2 Calibración lineal

Supongamos que hemos definido un modelo como los definidos anteriormente. Sea $L = \beta_0 + \beta_1 \cdot X_1 + ... + \beta_p \cdot X_p$. La calibración lineal consiste en ajustar un nuevo modelo de regresión logística, utilizando como variable dependiente Z y como variable independiente L. A través de éste, podemos calcular las probabilidades pronosticadas de nuestro evento y contrastarlas con las observadas. Esto se conoce como calibración lineal.[14-16]

$$logit(z) = a + b_L \cdot L.$$

3.1.3.2.3 Calibración por curvas suaves

Utilizando la notación anterior, para cada sujeto *i* definimos $l_i = {}_0 + {}_1 x_i$. A continuación, ajustamos un modelo de regresión logística binaria, empleando como variable explicativa una curva flexible dependiente del valor de *L*. Ésta puede construirse empleando curvas diferenciables definidas a trozos, como la transformación por splines o loess.[5] A través de este modelo estimamos las probabilidades de evento observadas, las cuales serán comparadas con las pronosticadas en un gráfico cartesiano. En éste podremos determinar si la curva de probabilidades (unión de todos los puntos) se ajusta a la línea diagonal (calibración perfecta).[14,17,18]

3.1.3.2.4 Medidas de la falta de calibración

Existen medidas escalares para determinar la falta de calibración que puede tener un modelo predictivo, es decir, cuánto se dispersan las probabilidades observadas de las esperadas. Aunque parece una buena medida para valorar un modelo predictivo, el hecho de obtener valores reducidos del mismo, no implica que el modelo sea correcto, es decir, hablamos de una condición necesaria pero no suficiente. Se destacan dos métodos para este propósito que veremos a continuación.[19,20]

3.1.3.2.4.1 Brier score.

El Brier score es la diferencia cuadrática media entre las observaciones de nuestra variable dependiente $Z(z_i)$ y las probabilidades pronosticadas $(\pi(x_1, ..., x_p))$. Es una medida global de la falta de calibración.[19,20]

3.1.3.2.4.2 Estimated Calibration Index.

El Estimated Calibration Index (ECI) es una medida que se ha propuesto para determinar la falta de calibración de un modelo predictivo. El ECI consiste en el cálculo de la diferencia media cuadrática entre el riesgo observado y el riesgo pronosticado por el modelo en un total de observaciones. El ECI tiene un rango de valores de 0 hasta 1, en donde el valor nulo corresponde a la absoluta perfección entre el modelo y la realidad. Aunque el ECI resume la falta de calibración en un único número, se ha observado que valores pequeños del mismo $(ECI = 1.67)$, producen modelos que no están bien calibrados. En otras palabras, si un modelo está bien calibrado, tendrá un valor de ECI reducido, pero esta implicación no ocurre a la inversa. En definitiva, es una condición necesaria, pero no suficiente. Consecuentemente

tendremos que representar el gráfico cartesiano de los modelos que obtengan un ECI pequeño.[20,21]

3.1.3.3 Evaluación del modelo

Tradicionalmente los estudios de validación de modelos de regresión logística se han realizado sobre una única muestra de sujetos, es decir, sobre esa única muestra de validación se ha calculado la discriminación y la calibración.[22] Sin embargo, en la actualidad la capacidad computacional de los ordenadores ha evolucionado y ello ha desembocado en la realización de la validación a través de muestras bootstrap.[5]

Dada una muestra de *n* sujetos, definimos muestra bootstrap como una muestra aleatoria con reposición de un total de *n* sujetos obtenida de la muestra original. Esta muestra aleatoria puede tomarse con varios diseños, como el muestreo aleatorio simple o el muestreo estratificado. En otras palabras, existirán elementos repetidos más de una vez y otros elementos que no pertenecerán a la muestra.[23]

La metodología del bootstrapping consiste en obtener un número elevado de muestras bootstrap, generalmente 1000 muestras, y en cada una de ellas determinar el valor de un parámetro estadístico y a través de estos valores construir la distribución de dicho parámetro. Si aplicamos esta metodología a la hora de validar un modelo de regresión logística, podremos construir la distribución del ABC, además de valorar la calibración.[23]

También existen otros métodos para la evaluación del modelo, como el cross-validation y el split-validation.[7,24-27] No obstante, la metodología del bootstrapping es la más recomendada para modelos predictivos.[5-7,10,22,24,28]

3.1.3.4 Tamaño muestral

La mayoría de estudios que se plantean en investigación clínica, como la estimación del AUC, tienen una fórmula cerrada para la obtención del tamaño muestral, la cual suele estar basada en una serie de parámetros (valores esperados poblacionales, error tipo I y II, razón entre muestras…). Sin embargo, la determinación de la calibración de un modelo predictivo no dispone de una fórmula cerrada.[21]

Por ese motivo se han realizado estudios de simulación para obtener cuántos pacientes necesitamos para poder afirmar con seguridad que el modelo predictivo está bien calibrado. Estos estudios han concluido que se necesitan al menos 100 eventos y 100 no eventos, independientemente del modelo predictivo que estemos abordando.[14,21,29,30]

3.1.4 Sistema de puntos

En las secciones anteriores se ha expresado la forma de calcular la probabilidad de un suceso mediante un modelo de regresión logística, conociendo una serie de variables independientes. La forma de cálculo es a través de una fórmula cerrada que incluye sumas, multiplicaciones y exponenciales. En consecuencia, sin la utilización de un dispositivo electrónico, no somos capaces de determinar esta probabilidad. Por ese motivo, los investigadores del Framingham Heart Study,[31] desarrollaron un algoritmo que adaptó estos modelos matemáticos para su utilización en la práctica clínica habitual sin la necesidad de emplear dispositivos electrónicos.[21]

4. Justificación

Artículo 1 (PLoS One)

A la hora de abordar el problema del cálculo del tamaño muestral para validar un modelo de predicción, existen factores que influyen en el gráfico de calibración, como son la parametrización del modelo, la incidencia del evento que se está valorando y la discriminación del modelo predictivo (ABC).[14,21] En otras palabras, no deberíamos de establecer un único valor (100 eventos y 100 no eventos) para comprobar la calibración de cualquier modelo predictivo.

Artículo 2 (Computer Methods and Programs in Biomedicine)

Para la utilización de un sistema de puntos de las características planteadas en una determinada población, es necesaria su validación.[5,6,22,26,32,33] La realización de dicha validación realizada con la metodología bootstrap no es una tarea sencilla,[1] por lo que es necesario dotar al profesional sanitario de una herramienta que determine todos los cálculos de validación de forma inmediata (aplicación para teléfono móvil). Dicha herramienta no ha sido implementada en la actualidad.

5. Hipótesis

Artículo 1 (PLoS One)

Puede plantearse la elaboración de un algoritmo que determine con mayor precisión el tamaño muestral necesario para validar un modelo de predicción y con ello disminuir el número de pacientes a recoger, sin que ello haga que la precisión en la determinación de la calibración disminuya.

Artículo 2 (Computer Methods and Programs in Biomedicine)

El software elaborado para el teléfono móvil en esta tesis doctoral podría tener gran relevancia clínica para la validación externa para sistemas de puntos basados en regresión logística e implementación posterior de los mismos en beneficio del paciente.

6. Objetivos

Artículo 1 (PLoS One)

1) Elaborar un algoritmo para determinar el número de sujetos para validar externamente un sistema de puntos basado en un modelo de regresión logística.

2) Aplicar el algoritmo sobre un sistema de puntos ya publicado, el cual valora la mortalidad en las unidades de cuidados intensivos (UCI).[34-36]

Artículo 2 (Computer Methods and Programs in Biomedicine)

1) Elaborar una metodología para construir una aplicación para teléfono móvil en el sistema Android en el que el profesional sanitario pueda validar a través de muestras bootstrap un sistema de puntos para predecir un evento binario y, una vez validado, sea capaz de aplicarlo sobre sus pacientes.

2) Aplicar la metodología a un caso práctico con datos simulados, con el objetivo de clarificar la metodología propuesta. Dicho ejemplo será el sistema de puntos de la UCI.[34-36]

3) Determinar en el caso práctico que no existen diferencias entre realizar la validación con el paquete estadístico R y con la aplicación móvil.

7. Material y métodos

Artículo 1 (PLoS One)

Algoritmo propuesto para calcular el tamaño muestral para validar externamente un sistema de puntos

Empleando todos los conceptos definidos anteriormente vamos a detallar un algoritmo para evaluar esta cuestión, ya que podría ser que necesitáramos un tamaño muestral con una cantidad de eventos y no eventos diferente de 100.[14,30] Hemos de tener en cuenta que se han de valorar dos aspectos (discriminación y calibración), de los cuales el primero de ellos no necesita un tamaño muestral excesivo para encontrar diferencias estadísticamente significativas.[30] Sin embargo, la obtención del tamaño muestral para determinar si un modelo está bien calibrado, requiere mayor estudio.[14] Por ese motivo, nos centraremos en el tamaño muestral para la calibración suave.

Vamos a comentar una serie de consideraciones: como se ha indicado anteriormente, el ECI es una medida que nos ayudará con esta tarea, ya que valores cercanos a 0 son condición necesaria, pero no suficiente, para decir que un modelo está bien calibrado. Por ello, estableceremos puntos de corte cercanos al valor nulo para el ECI y determinaremos si el modelo está bien calibrado a través de la interpretación del gráfico de calibración suave. Por otra parte, tenemos que tenemos que tener en cuenta que la variable aleatoria de las puntuaciones (x) puede considerarse con distribución multinomial, ya que tiene un número finito de valores. Además, dispondremos de la proporción de sujetos con evento (p_{event}). Finalmente, estableceremos un posible rango de valores para el número de eventos (n_{event}) con el objetivo de comprobar su calibración. Con estas consideraciones y los conceptos previamente analizados, pasamos a detallar el algoritmo propuesto:

1. Establecer n_{event} (si es la primera vez que se inicia este paso, n_{event} toma el valor mínimo del posible rango de valores para comprobar):

    a. Simular una muestra aleatoria del vector (x, z) a través de la distribución multinomial de las puntuaciones y del modelo de regresión logística asociado al sistema de puntos, con n_{event} sujetos con evento y con $n_{non-event} = \frac{1-p_{event}}{p_{event}} \cdot n_{event}$ sujetos sin evento. Notar que $n_{non-event}$ podría tener decimales, por lo que lo aproximaremos al número entero más cercano.

b. En la muestra del paso *1a*, determinar AUC y probabilidades observadas de evento para cada puntuación a través de curvas suaves.[14]

c. Repetir los pasos *1a* y *1b* un número prefijado de veces N (por ejemplo, 1000 veces) con el objetivo de construir la distribución de estos parámetros. Una vez repetidas las etapas anteriores veces, ir al paso *2*.

2. Determinar valor del ECI con el total de N observaciones realizadas, guardar el gráfico de calibración suave únicamente con sus intervalos de confianza y calcular los intervalos de confianza para el AUC. Notar que estamos interesados únicamente en los intervalos de confianza para tener un umbral de posibles valores poblacionales, es decir, asegurarnos con una probabilidad elevada que el modelo estará bien calibrado y que discriminará correctamente el sujeto con evento.

3. Recalcular n_{event} como $n_{event} = n_{event} + 1$ e ir al paso *1*, excepto si ya hemos comprobado todo el rango de posibles valores para , en cuyo caso iremos al paso *4*.

4. Con los puntos de corte determinados a priori para el ECI, crear variables indicadoras que determinen si con el número de eventos se verifica que el ECI es menor que dichos puntos de corte.

5. Construir las curvas ROC con n_{event} (variable cuantitativa) y las variables indicadoras del ECI menor que los puntos de corte.

6. Determinar el punto óptimo de n_{event} para cada uno de los puntos de corte del ECI, tal como se ha explicado en la subsección de las curvas ROC.

7. Interpretar los gráficos de calibración suaves de los tamaños muestrales obtenidos en el paso *6*.

8. Establecer como tamaño muestral el mínimo valor de los n_{event} del paso *6* que esté bien calibrado.

Aplicación a un caso real

Se va a aplicar el algoritmo propuesto sobre un sistema de puntos para predecir mortalidad en la UCI.[34-36] Los datos de dicho sistema son los siguientes:

Mínima puntuación: $x_{min} = 0$ puntos	*Máxima puntuación:* $x_{max} = 15$ puntos
Coeficientes del modelo de regresión logística	$\beta_0 = -5.92252114678228$ $\beta_1 = 0.6$
Proporción de eventos (p_{event})	0.10781990521327014218009478672986

Distribución de probabilidades asociadas a cada puntuación

x_0	0.29023508137432200	x_8	0.02441229656419530
x_1	0.03887884267631100	x_9	0.02622061482820980
x_2	0.09222423146473780	x_{10}	0.03345388788426760
x_3	0.18625678119349000	x_{11}	0.00994575045207957
x_4	0.05967450271247740	x_{12}	0.03526220614828210
x_5	0.08318264014466550	x_{13}	0.02893309222423150
x_6	0.06057866184448460	x_{14}	0
x_7	0.02893309222423150	x_{15}	0.00180831826401447

Estos datos han sido obtenidos de la publicación original.[34-36] En lo referente al rango de posibles valores para n_{event}, se ha establecido entre 25 y 1000. Las curvas suaves se han realizado mediante splines lineales.

Para visualizar la influencia del tamaño muestral sobre la discriminación y la calibración, se realizarán gráficos de líneas para los intervalos de confianza del AUC y para el ECI. Esta

evolución será analizada mediante un vídeo para los gráficos de calibración suave, en los que se apreciará el ajuste a la línea perfecta de las curvas conforme aumenta n_{event}. Los puntos de corte establecidos para el ECI han sido 2, 1.75, 1.5, 1.25, 1, 0.75, 0.5 y 0.25.

Artículo 2 (Computer Methods and Programs in Biomedicine)

Construcción de una aplicación móvil para Android

El código fuente se realiza en un lenguaje de programación, Java o Kotlin, en el entorno de desarrollo integrado Android Studio, donde se estructura en cuatro componentes principales: 1) *Setup*, 2) *Database*, 3) *Validation* y 4) *Predictor*. En primer lugar, *Setup* nos indica la situación inicial del sistema de puntos en la población del investigador, indicando si está validado y definiendo el tamaño muestral. La introducción de los datos de los pacientes se aborda en la siguiente componente (*Database*). En *Validation* aplicamos los métodos estadísticos para realizar la validación externa del sistema de puntos. Finalmente, si el sistema ha sido validado, ya sea porque el investigador lo ha realizado en un estudio previo o a través de la aplicación, *Predictor* lo aplica sobre un nuevo paciente determinando el riesgo del suceso binario de interés.

Ahora que hemos detallado la estructura, vamos a explicar el mecanismo de funcionamiento de la aplicación móvil a través de diagramas de flujo (Figs. 1-3). Lo primero que se nos plantea es indicarle a la aplicación si el sistema está validado (Fig. 1). En caso afirmativo, la aplicación nos permitirá el cálculo de la puntuación y la probabilidad de evento de un paciente. Por otro lado, si el sistema no está validado, deberemos calcular el tamaño muestral, el cual debe de incluir al menos 100 sujetos con evento y un número de sujetos sin evento proporcional al de la población. Esto significa que si en nuestra población esta proporción es $\widehat{p}$, deberíamos de tener al menos $\frac{1-\widehat{p}}{\widehat{p}} \cdot 100$ sujetos sin evento. No obstante, la cifra de 100 eventos es a nivel general para cualquier modelo predictivo sea cuales sean sus características, pero existe desde el año 2017 un algoritmo que determina el tamaño muestral para validar sistemas de puntos en particular,[21] indicando como ejemplo que para el sistema de puntos de la UCI con 69 eventos y $\frac{1-\widehat{p}}{\widehat{p}} \cdot 69$ pacientes sin evento, se disponía de un buen tamaño muestral. En la app podría darse al usuario la opción de elegir cualquiera de las dos aproximaciones. A continuación el usuario debe de introducir datos de pacientes hasta alcanzar este tamaño muestral, tras lo que la aplicación comenzaría el proceso de validación.

Al usuario se le ofrecerá una interfaz para la elaboración de la base de datos, donde éste podrá borrar, actualizar y añadir pacientes (Fig. 2). La base de datos, tras cálculo de la puntuación a través de las variables del sistema, tendrá las siguientes componentes: puntuación (x), indicador del evento (z), probabilidad de evento (p) y el predictor lineal (L), el cual será transformado en funciones splines. Observar que el valor de n (número total de sujetos introducidos hasta el momento en la base de datos) (Fig. 2), varía en función de

cada una de estas operaciones. Una vez hayamos introducido el tamaño muestral necesario, la aplicación móvil procederá a la validación del sistema de puntos empleando la metodología bootstrap (Fig. 3). Para facilitar esta tarea es recomendable establecer un contador de pacientes con evento y sin evento, además de proporcionar la lista de pacientes introducidos hasta el momento, para permitir al usuario editar o borrar registros.

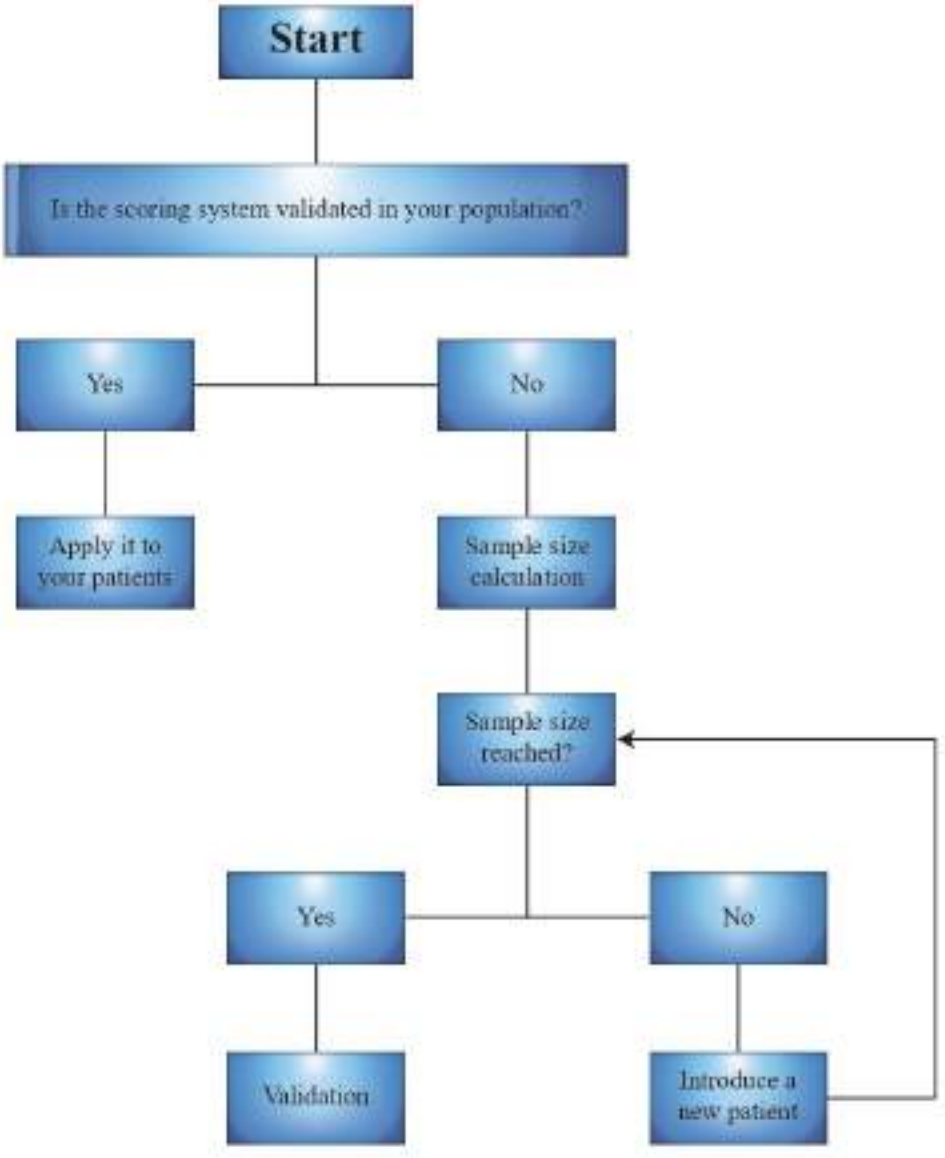


Fig. 1 Esquema general de la aplicación móvil.

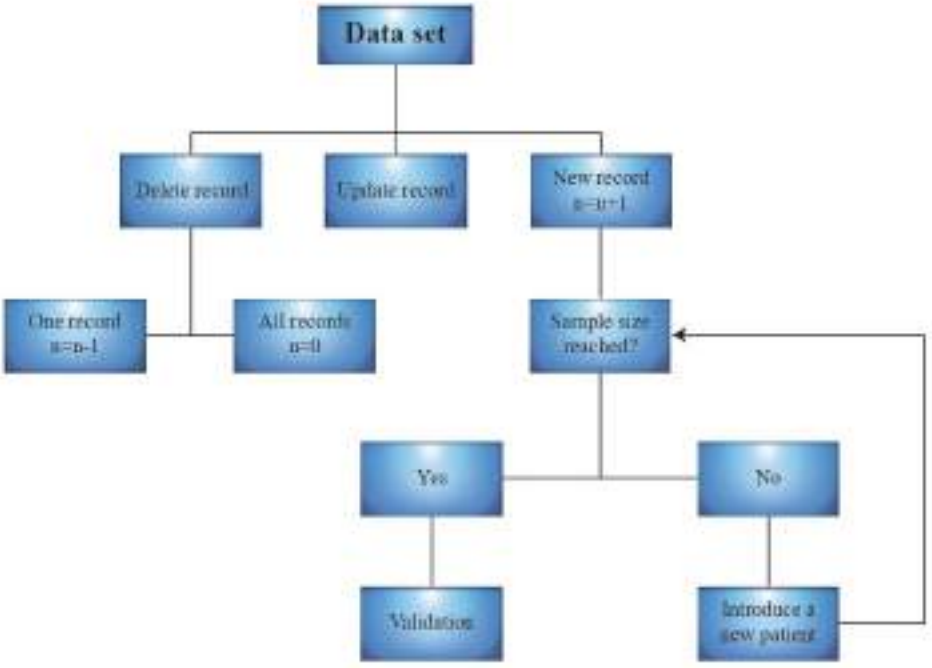


Fig. 2 Iteracción de la aplicación móvil con la base de datos. n, número de pacientes en la base de datos.

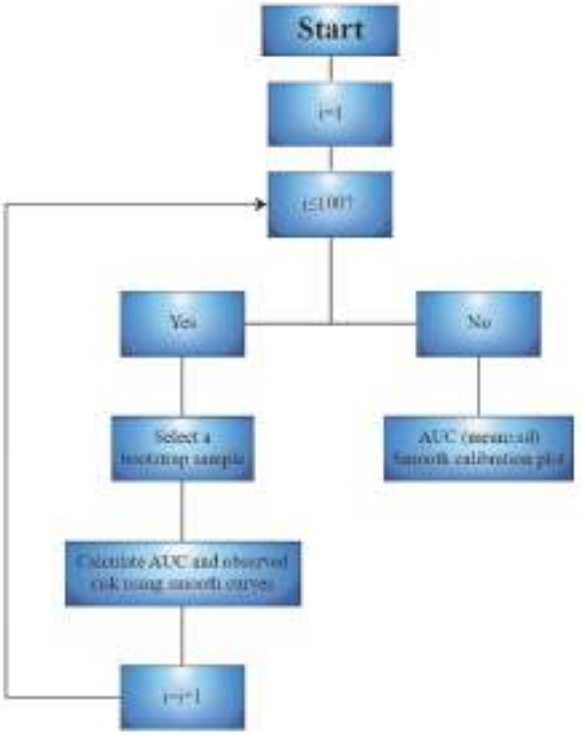


Fig. 3 Esquema general de la validación. AUC, área bajo la curva ROC; i, contador; sd, desviación estándar.

Para validar el sistema de puntos (Fig. 3), partimos de la base de datos introducida con n elementos, de una variable contador (i), de un vector vacío con 100 componentes para guardar los valores de la distribución del AUC, y de una matriz vacía con 100 filas y $x_{max} - x_{min} + 1$ columnas para guardar las probabilidades observadas para cada posible puntuación en cada muestra bootstrap. Esta variable contador (i) realiza un total de 100 iteraciones, donde en cada una de ellas obtiene una muestra bootstrap en la que calcula el AUC. En otras palabras, dado que se realiza 100 veces, se obtiene la distribución de este parámetro, la cual es utilizada para determinar si el sistema de puntos discrimina correctamente el evento en la población analizada. Como se ha mencionado anteriormente, el AUC tiene que tener un valor cercano a uno, para poder decir que discrimina correctamente qué paciente experimenta un evento. Esto se valorará a través del cálculo de la media y de la desviación estándar. Por otro lado, la aplicación determina en cada muestra bootstrap las probabilidades observadas de evento para cada puntuación del sistema a través de curvas suaves, con las cuales se construye el gráfico de calibración, el cual debe de valorarse por el usuario en función del ajuste que tenga la curva a la línea diagonal (observado=esperado). Un punto a tener en cuenta es que la calibración suave podría no converger en algunos sistemas de puntos, por lo que se debería de aplicar otro método de calibración, como es la lineal. No obstante, se debería de indicar unas guías para poder interpretar los resultados, en función de la dispersión existente entre lo observado y lo esperado. Esta información, junto con la base de datos introducida, debería de ser enviada al usuario de la aplicación mediante un email o método similar.

En caso de que hayamos obtenido resultados satisfactorios para poder concluir que el sistema ha sido validado externamente en nuestra población, al igual que si al arrancar la propia aplicación hemos indicado que el sistema estaba previamente validado, ésta tendría que permitir al usuario aplicar el sistema sobre un nuevo paciente, introduciendo las variables del mismo y determinando el riesgo del outcome.

Aplicación a un caso real (ICU mortality)

Para facilitar al lector la comprensión del algoritmo para desarrollar la aplicación móvil, se ha aplicado al sistema de puntos para predecir mortalidad en la UCI.[34-36] Este sistema otorga una puntuación total a cada sujeto, basada en la suma de las puntuaciones parciales de las variables: ingreso médico, sepsis, soporte inotrópico, ingreso cardiológico, ventilación mecánica y escala funcional (independiente, dependiente and discapacitado).[34-36] La aplicación móvil ha sido subida a Google Play con el nombre de *ICU mortality* y su descarga es libre y gratuita para cualquier usuario. La aplicación realiza la calibración suave a través de splines transformations, salvo que no converja que realizará la calibración lineal. En la Fig. 4 se aprecian capturas de las diferentes fases de la mobile application.

Fig. 4 Capturas de pantalla de la aplicación. A, cálculo de la muestra ;B, introducción nuevo paciente; C, informe de la validación; D, informe del sistema puntos después de la validación.

Respecto a la utilización de la aplicación para validar el sistema de puntos planteado, con el único objeto de que se pueda visualizar todo el proceso sin introducir ningún dato, hemos simulado una base de datos similar a aquélla sobre la que fue desarrollado el sistema de puntos. Dicha simulación ha sido basada en modelos de regresión logística cuyos coeficientes han sido obtenidos con la base de datos original.[34-36] Para que el lector sea capaz de replicar nuestro conjunto de datos, se ha incorporado como material suplementario de la publicación asociada al método,[37] tanto el código en R correspondiente como la base de datos generado por dicho código y utilizada para mostrar los resultados obtenidos. La muestra simulada (fixtures) se cargará en la aplicación pulsando 7 veces sobre el contador que muestra cuántos pacientes se han introducido y tendrá el tamaño muestral necesario para validar a excepción de un paciente fallecido, para que el usuario pueda introducirlo manualmente y vea el proceso completo. Finalmente, indicar que la aplicación calcula la matriz de splines de grado 1 (B-spline basis matrix), utilizando como nodos los percentiles 20, 40, 60 y 80.

Comparación entre los resultados obtenidos con el paquete estadístico y la aplicación móvil

Las distribuciones obtenidas en ambos dispositivos (R y aplicación móvil) para nuestros parámetros (AUC y probabilidades suaves para cada puntuación) fueron comparadas a nivel descriptivo y gráfico.

8. Resultados

Artículo 1 (PLoS One)

En la Fig. 5 se observa la evolución del AUC conforme aumenta el número de eventos en la muestra, mientras que en la Fig. 6 esta misma evolución es representada para el ECI.

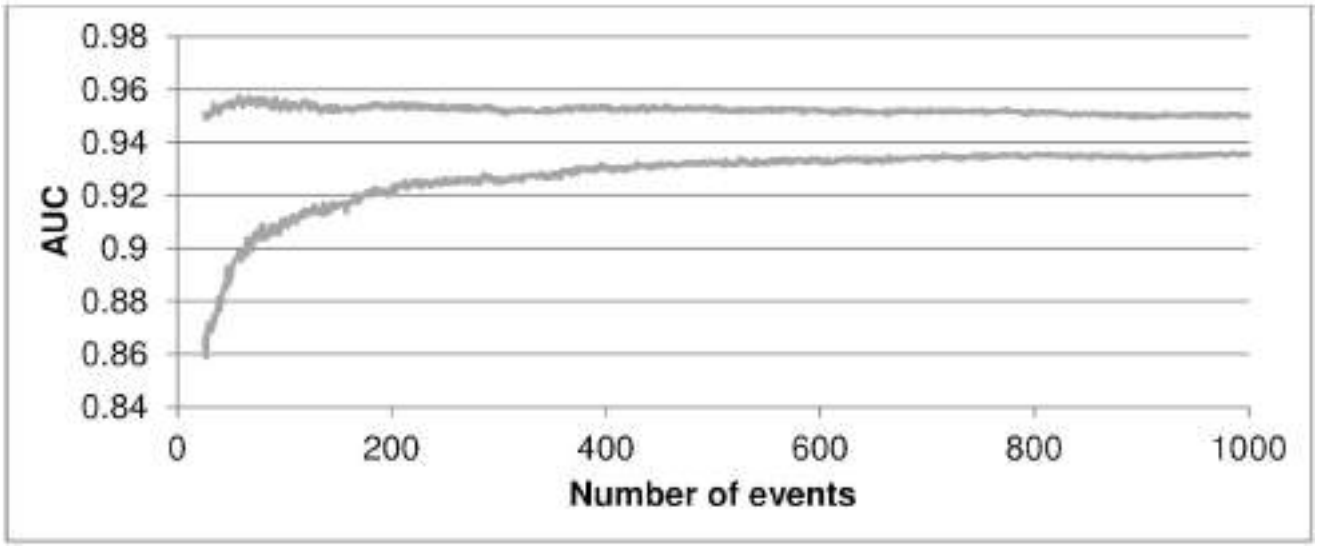


Fig. 5 Intervalos de confianza del AUC en función del número de eventos en la muestra. AUC, área bajo la curva ROC.

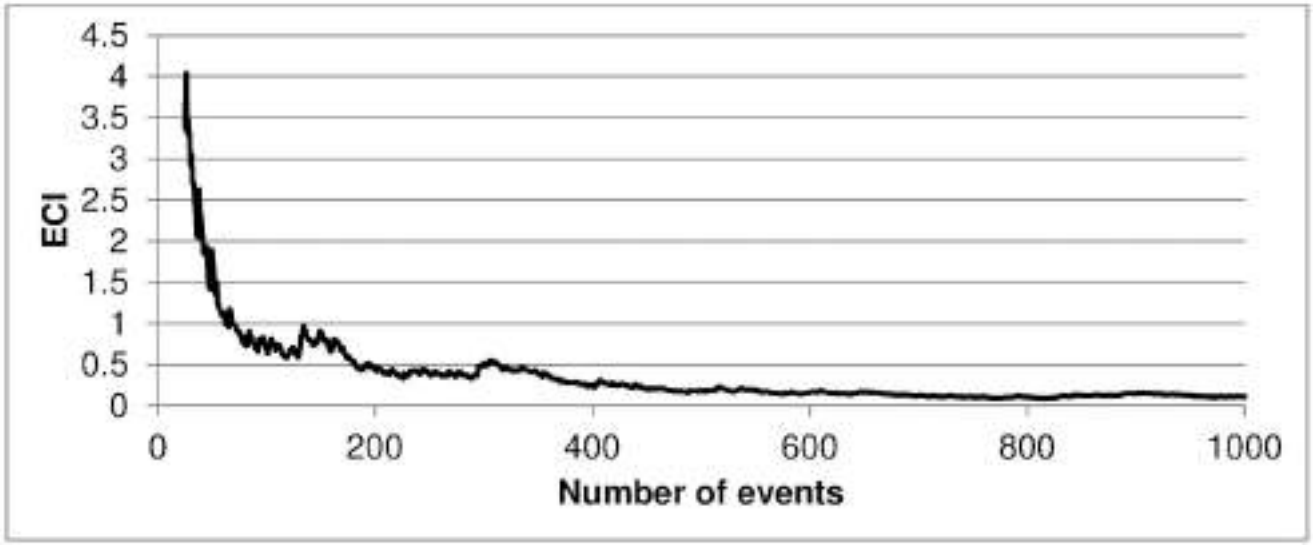


Fig. 6 Valores del ECI en función del número de eventos en la muestra. ECI, Estimated Calibration Index.

Con los puntos de corte escogidos para el ECI (2, 1.75, 1.5, 1.25, 1, 0.75, 0.5 y 0.25), el número de eventos para la muestra indicados siguiendo nuestro algoritmo fueron 42, 55, 56, 69, 167, 196 y 430, respectivamente. Los gráficos de calibración suaves para dichos tamaños, junto con el valor inicial del rango comprobado ($n_{event} = 25$), el valor sugerido en la bibliografía ($n_{event} = 100$) y el valor final de dicho rango ($n_{event} = 1000$), han sido representados en la Fig. 7. En ella vemos que el mínimo número de eventos con buena calibración es 69, complementado con un ECI<1.25. Si calculamos el tamaño muestral a través de p_{event}, éste es de 640 pacientes (69 pacientes fallecidos y 571 vivos). Este tamaño sería de 100 pacientes fallecidos y de 828 vivos (938 en total) según las recomendaciones de la bibliografía, lo que supone 298 pacientes más que siguiendo nuestro algoritmo.

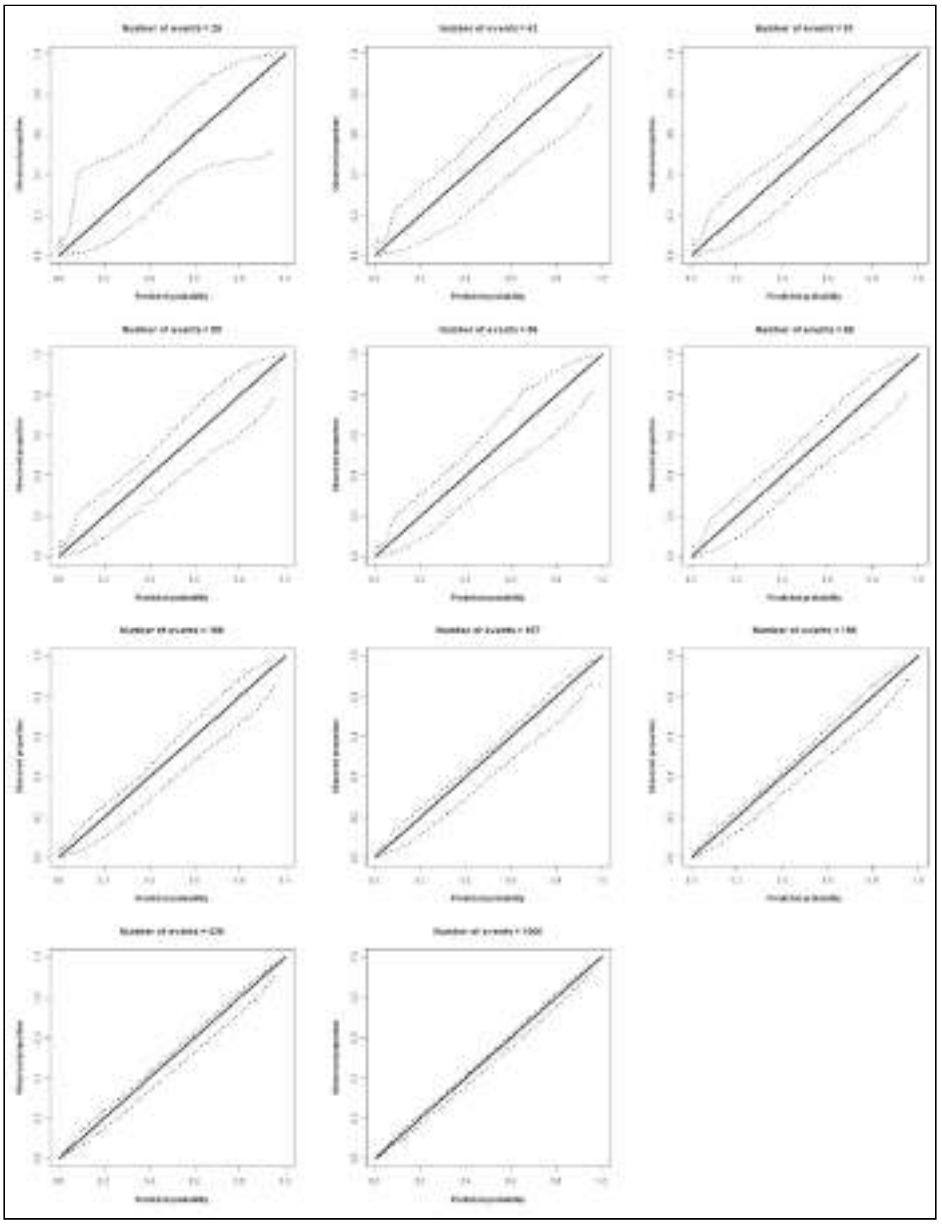


Fig. 7 Gráficas de calibración por curvas suaves (splines lineales) para varios tamaños muestrales. Las líneas discontinuas son los intervalos de confianza y la línea central denota la condición perfecta.

Artículo 2 (Computer Methods and Programs in Biomedicine)

A la hora de comparar los resultados obtenidos con el paquete estadístico R y con la mobile application, no se apreciaron apenas diferencias ni para el AUC (R, 0.94±0.01; aplicación móvil, 0.93±0.01) ni para la calibración (Fig. 8).

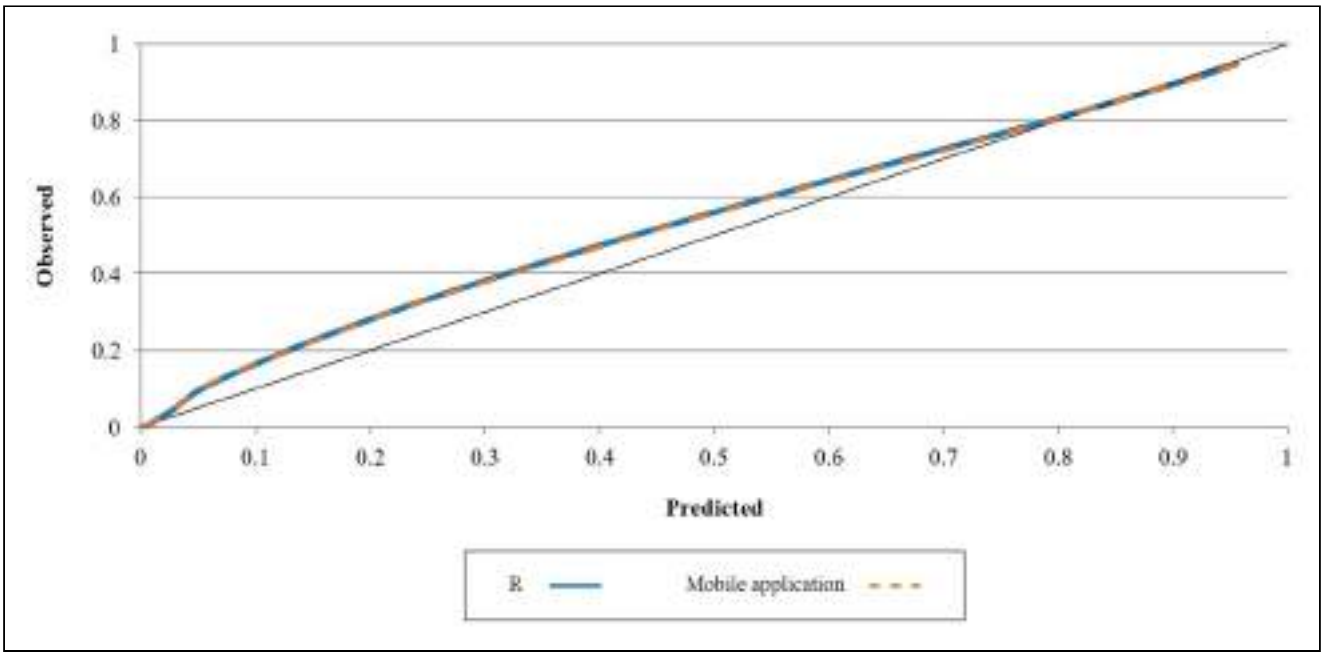


Fig. 8 Comparativa de la calibración entre el paquete estadístico R y nuestra aplicación móvil. La línea fina denota la condición perfecta.

En primer lugar, los valores del AUC difieren sólo en una centésima parte, lo cual es irrelevante y podría deberse a las muestras de bootstrap seleccionadas o incluso al redondeo de los propios valores para tener dos cifras significativas. En segundo y último lugar, las curvas suaves de los dos procedimientos se superponen claramente, lo que demuestra que los valores de las probabilidades observadas son muy similares. Todo esto era lógico y esperado, ya que los métodos matemáticos utilizados eran los mismos en estos procedimientos (R y nuestra aplicación móvil).

9. Discusión

En esta tesis doctoral se ha detallado un algoritmo para que cualquier lector pueda programar una aplicación para teléfono móvil en sistema Android, para poder validar cualquier sistema de puntos derivado de un modelo de regresión logística para predecir un suceso binario. Previamente, con el objetivo de optimizar el proceso de validación, se ha desarrollado un método para determinar el tamaño muestral para dicho proceso. Ambas técnicas han sido aplicadas sobre datos simulados, de forma que se comprenda con mayor claridad las distintas etapas de las mismas.

9.1 Fortalezas y limitaciones

Artículo 1 (PLoS One)

La fortaleza principal de este trabajo es la idea desarrollada, ya que por primera vez se dispone de un algoritmo para calcular el tamaño muestral para validar externamente sistemas de puntos basados en modelos de regresión logística binaria. Esta cuestión no ha sido abordada en profundidad en la literatura científica, ya que se ha recomendado la utilización de 100 eventos y 100 no eventos, independientemente de las características del modelo.[14,30] Este valor de 100 no debería de ser fijo, pues se ha visto que hay factores que pueden influir en el gráfico de calibración.[18] En segundo lugar, destacamos la utilización de curvas suaves en vez de categorizaciones de riesgo, como por ejemplo el test de Hosmer-Lemeshow, pues dan mayor validez a los resultados.[14] Finalmente, pensamos que este algoritmo puede extenderse a casos más complejos, como sistemas de puntos basados en modelos de supervivencia, o modelos de regresión logística/supervivencia en general, ya que los sistemas de puntos son un caso particular de los mismos.

Como limitación indicamos que este cálculo requiere un coste computacional elevado, ya que son necesarias múltiples simulaciones para cada número de eventos del rango propuesto. En nuestro caso nuestro rango tenía 976 posibles valores y 1000 simulaciones en cada valor, lo que equivale a un total de 976,000 simulaciones, en las cuales se calcula el AUC y los valores observados a través de curvas suaves. No obstante, si tenemos en cuenta el beneficio que nos puede aportar el empleo de este algoritmo, esto no supondría una limitación. En el ejemplo propuesto hemos reducido en 298 pacientes el tamaño muestral sugerido por la literatura, lo que equivale a la reducción considerable de costes, tanto económicos como del tiempo necesario para reclutar a los participantes del estudio. En otras palabras, el algoritmo presenta utilidad para valorar la cuestión a estudio (tamaño muestral para validar sistemas de puntos basados en modelos de regresión logística binaria).

Artículo 2 (Computer Methods and Programs in Biomedicine)

Uno de los puntos que da mayor fortaleza a este trabajo, es la elaboración de un algoritmo para que se puedan validar sistemas de puntos para predecir eventos binarios basados en modelos de regresión logística automáticamente, simplemente con un teléfono con sistema operativo Android, lo que permite que dichos sistemas, tras utilización de nuestra aplicación móvil, puedan implementarse en la práctica clínica habitual de forma inmediata. Como limitación se indica que en la validación a través de bootstrap, al tener que realizar el teléfono multitud de cálculos matemáticos, es posible que el usuario no pueda utilizar su teléfono para otras funciones mientras se realiza la validación. No obstante, dado que el procesador de los teléfonos móviles cada día es más potente, esta cifra irá reduciéndose de forma acelerada con la implementación de nuevos procesadores en la telefonía móvil. Por otra parte, si tenemos en cuenta el tiempo que se puede tardar en introducir las características de los sujetos en una base de datos en un ordenador y luego aplicar técnicas de validación a través de un paquete estadístico, nuestros minutos de utilización de la aplicación móvil, pueden suponer un pequeño porcentaje de dicho tiempo.

9.2 Comparación con la literatura existente

Artículo 1 (PLoS One)

A la hora de comparar nuestro algoritmo para optimizar el tamaño muestral con aquello publicado en la literatura científica, observamos que los otros estudios no han considerado las características particulares de cada modelo predictivo en sí, es decir, que han dado un tamaño muestral general para la validación de cualquier modelo predictivo.[14,30] Como se ha mencionado anteriormente, esto no es del todo correcto, ya que en función del modelo predictivo que estemos validando externamente, el tamaño muestral para dicha validación debería de ser independiente.[18] No obstante, hemos de tener en cuenta que el algoritmo que hemos desarrollado es para sistemas de puntos, los cuales son un caso particular de los modelos de regresión logística binaria.[31]

En lo referente a los puntos de corte del ECI, hemos establecido este sistema para determinar nuestro tamaño muestral. No obstante, otra forma de abordar este problema, podría ser analizar todos los gráficos de calibración y determinar visualmente cuál es el primero de ellos en el que podemos decir que el sistema de puntos está bien calibrado. Nosotros hemos querido incorporar el ECI, ya que es una forma objetiva de medir la falta de calibración, la cual complementada con el gráfico de calibración, hace que podamos ver de forma más rigurosa la cuestión analizada.[14]

Artículo 2 (Computer Methods and Programs in Biomedicine)

Hasta donde nosotros conocemos, no se han desarrollado algoritmos similares.

9.3 Implicaciones para la investigación

Artículo 1 (PLoS One)

Como nueva línea de investigación proponemos adaptar este algoritmo a sistemas de puntos basados en modelos de supervivencia. Para ello deberemos de fijar puntos de corte para el tiempo de predicción y obtener las probabilidades observadas de evento en dichos puntos de corte a través de curvas suaves. Estas probabilidades dependerán del correspondiente valor en el sistema de puntos y de la supervivencia basal en el instante de tiempo que estemos valorando.[31] Por otro lado, animamos a otros autores a adaptar nuestro algoritmo a modelos de regresión logística en general.

Respecto a la utilización de nuestro algoritmo tal cual ha sido descrito, recomendamos su utilización para calcular el tamaño muestral para validar externamente sistemas de puntos basados en modelos de regresión logística binaria. De esta forma sabremos cuántos pacientes requerimos para dicho estudio, ya que es posible que necesitemos menos (o más) de 100 eventos y 100 no eventos, tal como se ha determinado en la literatura científica.[14,30]

De acuerdo con los resultados obtenidos en el caso práctico realizado (mortalidad en ICU), se ha obtenido un tamaño muestral de 640 pacientes, en los cuales se incluyen 69 fallecidos. En consecuencia si otros autores desean realizar estudios para validar externamente el sistema de puntos para predecir mortalidad en ICU,[34-36] disponen del cálculo del tamaño muestral para dichos estudios.

Artículo 2 (Computer Methods and Programs in Biomedicine)

Podría decirse que la metodología para construir nuestra aplicación móvil necesita de validación, pero si tenemos en cuenta que las técnicas matemáticas empleadas han sido utilizadas ampliamente en la literatura científica para validar sistemas de puntos y que nuestro algoritmo se ciñe a dichas técnicas, no es necesario dicho estudio de validación.[21] En otras palabras, disponemos una metodología que puede ser aplicada para cualquier sistema de puntos para predecir un suceso binario a través de un modelo de regresión logística, la cual se implementa en una aplicación móvil, lo que consigue que cualquier usuario valide un sistema de puntos y en caso de obtener resultados satisfactorios en dicha validación, éste implemente el sistema para su práctica clínica habitual. Finalmente, al estar la aplicación disponible para cualquier usuario (Google Play: ICU mortality), animamos a otros investigadores a validar este sistema de puntos en sus poblaciones, al igual que a implementar nuestra metodología en otros sistemas de puntos.

10. Conclusión

Artículo 1 (PLoS One)

1) Se ha construido un algoritmo para determinar el tamaño muestral para validar sistemas de puntos basados en modelos de regresión logística binaria., basándose en conceptos básicos a la hora de validar un modelo predictivo (curva ROC, gráficos de calibración suave y ECI).

2) Hemos aplicado el algoritmo sobre un caso real para ayudar a comprender mejor su aplicación.

Artículo 2 (Computer Methods and Programs in Biomedicine)

1) Se ha detallado una metodología para construir una aplicación móvil para el sistema operativo Android, que valida sistemas de puntos basados en modelos de regresión logística para predecir eventos binarios a través de bootstrap. Dicha validación está basada en técnicas matemáticas ampliamente utilizadas en la literatura científica.

2) Para facilitar la comprensión de nuestra metodología, se ha incorporado un ejemplo práctico y en dicho ejemplo se obtienen los resultados de validación, tanto a través de la aplicación móvil como por el paquete estadístico R.

3) No se han encontrado diferencias en el caso práctico entre realizar la validación con el paquete estadístico R y con la aplicación móvil.

11. Referencias

1. Hosmer D, Lemeshow S. Applied Logistic Regression. 2nd Ed. New York, NY, United States of America: Wiley; 2000.

2. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ* 2020; **369**: m1328.

3. Hauck WW, Donner A. Wald's Test as applied to hypotheses in logit analysis. *J Am Stat Assoc* 1977; **72**: 851–3.

4. Jennings DE. Judging inference adequacy in logistic regression. *J Am Stat Assoc* 1986; **81**: 471–6.

5. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014; **11**: e1001744.

6. Harrell FE. Regression Modeling Strategies. New York, NY, United States of America: Springer-Verlag; 2001.

7. Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and updating. Rotterdam, Netherlands: Springer; 2009: 497 p.

8. 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**: 361–87.

9. Royston P, Altman DG. External validation of a cox prognostic model: principles and methods. *BMC Med Res Methodol* 2013; **13**: 33.

10. Moons KG, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012; **98**: 683–90.

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

12. Steyerberg EW, Eijkemans MJ, Harrell FE, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making* 2001; **21**: 45–56.

13. Peek N, Arts DG, Bosman RJ, van der Voort PH, de Keizer NF. External validation of prognostic models for critically ill patients required substantial sample sizes. *J Clin Epidemiol* 2007; **60**: 491–501.

14. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016; **74**: 167–76.

15. Cox DR. Two further applications of a model for binary regression. Biometrika 1958; **45**: 562–5.

16. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. *Med Decis Making* 1993; **13**: 49–58.

17. Harrell Jr FE. Regression modeling strategies. With applications to linear models, logistic regression, and survival analysis. New York, NY, United States of America: Springer-Verlag; 2001.

18. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med* 2014; **33**: 517–35.

19. Van Hoorde K, Van Huffel S, Timmerman D, Bourne T, Van Calster B. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. *J Biomed Inform* 2015; **54**: 283–93.

20. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**: 128–38.

21. Palazón-Bru A, Folgado-de la Rosa DM, Cortés-Castell E, López-Cascales MT, Gil-Guillén VF. Sample size calculation to externally validate scoring systems based on logistic regression models. *PLoS One* 2017; **12**: e0176726.

22. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000; **19**: 453–73.

23. Efron B. Bootstrap methods: Another look at the jackknife. *Ann Stat* 1979; **7**: 1–26.

24. Steyerberg EW, Moons KG, van der Windt DA, et al.Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013; **10**: e1001381.

25. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. *J Clin Epidemiol* 2010; **63**: e1–e37.

26. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009; **338**: 1432–5.

27. Jacob M, Bruegger D, Conzen P, et al. Development and validation of a mathematical algorithm for quantifying preoperative blood volume by means of the decrease in hematocrit resulting from acute normovolemic hemodilution. *Transfusion* 2005; **45**: 562–71.

28. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ* 2009; **338**: b604.

29. 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**: 475–83.

30. 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**: 214–26.

31. Sullivan LM, Massaro JM, D'Agostino RB Sr. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med* 2004; **23**: 1631–60.

32. 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.

33. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012; **98**: 691–8.

34. Dólera-Moreno C, Palazón-Bru A, Colomina-Climent F, Gil-Guillén VF. Construction and internal validation of a new mortality risk score for patients admitted to the intensive care unit. *Int J Clin Pract* 2016; **70**: 916–22.

35. Palazón-Bru A, Dólera-Moreno C, Folgado-de la Rosa DM, Colomina-Climent F, Gil-Guillén VF. An update to the internal validation of the new mortality risk score for patients admitted to the intensive care unit. *Int J Clin Pract* 2016; **70**: 961–2.

36. Palazón-Bru A, Colomina-Climent F, Dólera-Moreno C, Folgado-de la Rosa DM, Gil-Guillén VF. A brief comment about predictive models for mortality in intensive care units. *Acta Anaesthesiol Scand* 2018; **62**: 404.

37. Folgado-de la Rosa DM, Palazón-Bru A, Gil-Guillén VF. A method to validate scoring systems based on logistic regression models to predict binary outcomes via a mobile application for Android with an example of a real case. *Comput Methods Programs Biomed* 2020; **196**: 105570.

12. Anexos

En esta sección se incluyen los dos artículos científicos que forman parte de la producción científica de esta tesis doctoral. A continuación se detallan los indícios de calidad de cada una de las publicaciones:

Artículo 1 (PLoS One)

Según datos del *Journal Citations Reports* del año 2017, la revista PLoS One, perteneciente a la editorial Public Library of Science (PLOS), tenía un factor de impacto de 2.766, lo cual la situaba en el primer cuartil del área de Multidisciplinary Sciences (posición 15 de un total de 64).

Con respecto al número de citaciones del artículo a día de hoy (30 de octubre de 2020), según Google Scholar, el trabajo ha recibido un total de 22 en un periodo aproximado de 3 años y medio. Esto equivale a una media de 6 referencias recibidas cada año.

Artículo 2 (Computer Methods and Programs in Biomedicine)

El artículo fue publicado en noviembre del año 2020 en una revista, que según datos del *Journal Citation Reports* del año 2019 (todavía no disponible la última actualización), tenía un factor de impacto de 3.632, lo que la sitúa en el percentil 85 (posición 16 de 108) en el área Computer Science, Theory & Methods. En consecuencia, se sitúa al igual que la anterior publicación, en el primer cuartil.

Con respecto al número de citas, al ser un artículo muy reciente, no ha recibido en la actualidad ninguna referencia.

PLOS ONE

RESEARCH ARTICLE

Sample size calculation to externally validate scoring systems based on logistic regression models

Antonio Palazón-Bru[1☯]*, David Manuel Folgado-de la Rosa[1☯], Ernesto Cortés-Castell[2], María Teresa López-Cascales[3], Vicente Francisco Gil-Guillén[1]

1 Department of Clinical Medicine, Miguel Hernández University, San Juan de Alicante, Alicante, Spain, **2** Department of Pharmacology, Pediatrics and Organic Chemistry, Miguel Hernández University, San Juan de Alicante, Alicante, Spain, **3** Department of Molecular Neurobiology, Neurosciences Institute (Miguel Hernández University and Consejo Superior de Investigaciones Científicas), San Juan de Alicante, Alicante, Spain

☯ These authors contributed equally to this work.
* antonio.pb23@gmail.com

OPEN ACCESS

Citation: Palazón-Bru A, Folgado-de la Rosa DM, Cortés-Castell E, López-Cascales MT, Gil-Guillén VF (2017) Sample size calculation to externally validate scoring systems based on logistic regression models. PLoS ONE 12(5): e0176726. https://doi.org/10.1371/journal.pone.0176726

Editor: Ali Montazeri, Iranian Institute for Health Sciences Research, ISLAMIC REPUBLIC OF IRAN

Received: November 2, 2016

Accepted: April 14, 2017

Published: May 1, 2017

Data Availability Statement: All relevant data are within the paper (simulation algorithm).

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Background

A sample size containing at least 100 events and 100 non-events has been suggested to validate a predictive model, regardless of the model being validated and that certain factors can influence calibration of the predictive model (discrimination, parameterization and incidence). Scoring systems based on binary logistic regression models are a specific type of predictive model.

Objective

The aim of this study was to develop an algorithm to determine the sample size for validating a scoring system based on a binary logistic regression model and to apply it to a case study.

Methods

The algorithm was based on bootstrap samples in which the area under the ROC curve, the observed event probabilities through smooth curves, and a measure to determine the lack of calibration (estimated calibration index) were calculated. To illustrate its use for interested researchers, the algorithm was applied to a scoring system, based on a binary logistic regression model, to determine mortality in intensive care units.

Results

In the case study provided, the algorithm obtained a sample size with 69 events, which is lower than the value suggested in the literature.

Conclusion

An algorithm is provided for finding the appropriate sample size to validate scoring systems based on binary logistic regression models. This could be applied to determine the sample size in other similar cases.

Introduction

The predictive model most widely used in medicine to determine the onset of a clinical event (disease, relapse, death, healing. . .) is the binary logistic regression model. The probability of an event based on a series of parameters (explanatory variables) is obtained through a closed formula including addition, multiplication and exponentials [1]. Consequently, we are unable to determine this probability without the use of an electronic device. For this reason, researchers from the Framingham Heart Study developed an algorithm that adapted these mathematical models for use in routine clinical practice without the need for electronic devices, using scoring systems [2].

The algorithm begins by categorizing all the explanatory variables, associating each category with a score obtained through weighting the model coefficients. This then gives a finite set of total scores which: 1) is determined by the sum of all the scores associated with each of the explanatory variables, and 2) has an associated event probability [2]. In other words, the algorithm transforms a multivariate binary logistic regression model into another with a single explanatory variable (total score), which has a finite number of values, allowing the event probability to be calculated previously for each score.

Both the logistic regression models and the particular case of scoring systems must be validated externally for use in other populations. To carry out this process, both discrimination and calibration must be examined [3]. Discrimination consists of determining whether a higher event probability predicted by the model can differentiate between those subjects who experience an event and those who do not. To address this question, the area under the receiver operating characteristic (ROC) curve (AUC) is calculated [4]. Calibration involves analyzing whether the event probabilities predicted by the model correspond to those observed in reality. Generally, this process has been evaluated by categorizing into risk groups and through the logistic recalibration framework with a linear predictor [5]. However, it is preferable and advisable to use smooth calibration plots based on linear splines or loess [5,6].

When any study requiring statistical tests is performed, such as the external validation of a predictive model, it is necessary to calculate the number of subjects needed to accurately conclude that the results obtained in the sample can be extrapolated to the study population [7]. Most studies undertaken in clinical research, such as estimating the AUC [4], have a closed formula for obtaining the sample size based on a set of parameters (expected population values, type I and type II error, ratio between samples. . .). However, the determination of the calibration of a predictive model does not have a closed formula. For this reason, simulation studies have been performed to ascertain how many patients are needed to be able to say that the predictive model is well calibrated [5,8]. These studies have concluded that it takes at least 100 events and 100 non-events, regardless of the predictive model being addressed [5,8]. However, when approaching the problem of calculating sample size, factors exist that influence the calibration plot, such as model parameterization [5], incidence of the event being assessed, and the discrimination of the predictive model (AUC) [9]. In other words, we should not establish a single value (100 events and 100 non-events) to check the calibration of all predictive models.

Considering the usefulness of scoring systems in medicine (concrete case of logistic regression models) the fact that there is just one single sample size to validate any predictive model (despite the influence of different factors and that data collection may be laborious) means it is necessary to optimize the sample size so that it can efficiently validate a scoring system statistically without having to collect an excessive number of patients.

The objective of this paper is to explain an algorithm to determine the number of subjects to externally validate a scoring system based on a logistic regression model, which is a particular type of predictive model with a single linear predictor. In other words, we are determining the sample size calculation to externally validate a scoring system of the detailed characteristics. To illustrate how to use this algorithm, it will be applied to an already published scoring system that assesses mortality in intensive care units (ICU) [10]. To address these issues we will adhere to the following structure: first, a synthesis of the concepts of calibration by smooth curves and the AUC, followed by details of the suggested algorithm (sample size calculation). This algorithm will then be applied to the scoring system for mortality in the ICU and finally, a methodological discussion of the proposed algorithm will be provided.

Materials and methods

The area under the receiver operating characteristic curve

Suppose we have a random sample of n subjects {1,2,. . .,i,. . .,n}, where for each subject we have collected two random variables x_i and z_i, where x is a quantitative variable (discrete or continuous) and z an event indicator variable, i.e., it takes the value 1 when a subject has experienced the event and 0 when a subject has not. Our goal is to determine whether the variable x can discriminate (differentiate or distinguish) between subjects who experience an event and those who do not; that is, if higher values of x are associated with an increased event probability. Note that this could be done in the opposite way; i.e., smaller values of x associated with an increased event risk. Without loss of generality, we will proceed using the first method, as we can move from the second to the first case by multiplying the variable x by −1.

We define the sets: $E = \{i: z_i = 1, i = 1,. . .,n\}$ and $\bar{E} = \{i \ : \ z_i = 0, i = 1, \ldots, n\}$, equivalent to subjects who have experienced an event and those who have not, respectively. Note that $E \cup \bar{E} = \{1, \ldots, n\}$ and $E \cap \bar{E} = \emptyset$. With all these elements we are able to define the ROC curve [4], which is obtained by joining the following points on a Cartesian graph restricted to [0,1] x [0,1]:

$$\left(1 - \frac{|i \in \bar{E} \ : \ x_i < x|}{|\bar{E}|}, \frac{|i \in E \ : \ x_i \geq x|}{|E|}\right) x \in \{x_1, x_2, \ldots, x_i, \ldots, x_n\},$$

with $|\cdot|$ being the cardinal function of a given set, i.e., the number of elements contained in said set. For any value $\tilde{x}$ of the random variable x, the two components of each point on the Cartesian graph correspond respectively to 1-specificity and the sensitivity of a diagnostic test in which positive is defined as $x \geq \tilde{x}$ and negative as $x < \tilde{x}$ [11].

To calculate the area under the curve in the space [0,1] x [0,1] (AUC), assume two subjects $j \in E$ and $l \in \bar{E}$. Now we define:

$$S(j,l) = \begin{cases} 1 & if \ x_j > x_l \\ 1/2 & if \ x_j = x_l \\ 0 & if \ x_j < x_l \end{cases}.$$

The calculation of the AUC is obtained through [4]:

$$AUC = \frac{1}{|E| \cdot |\bar{E}|} \cdot \sum_{j \in E} \sum_{l \in \bar{E}} S(j,l).$$

Note that if x is a continuous variable $S(j,l)$ it will never take the value of 1/2.

The AUC is a way to measure the discrimination of a quantitative variable regarding the occurrence of an event. Its interpretation is the following: the closer the AUC is to one indicates that the variable x discriminates to a higher degree which subject has experienced an event [4].

We are now interested in determining an $\hat{x}$ value of the variable x (cut-off point) to distinguish with minimal error between subjects with and without an event, i.e., consider positive (subject with event) if $x \geq \hat{x}$ and negative in the opposite case (subject without event). The literature on ROC curves uses that value of the random variable x that minimizes $\sqrt{(1 - Sensitivity(x))^2 + (1 - Specificity(x))^2}$ [11].

Scoring systems based on logistic regression models

A scoring system is defined by the following elements [2]: 1) A set of possible score values (consecutive integers): $\{x_{min}, x_{min}+1, \ldots, -1, 0, 1 \ldots, x_{max}-1, x_{max}\}$, where x_{min} and x_{max} represent the minimum and maximum score of the system, respectively. We now denote x as the score variable, which has $x_{max} - x_{min} + 1$ possible values.

2) A binary logistic regression model defined as $logit(z) = \beta_0 + \beta_1 \cdot x$, with z being the indicator variable of the event and β_0 and β_1 the model coefficients associated with the constant and the varying score, respectively. Through these parameters (β_0 and β_1) we can obtain the random variable event probability p for each score x by the expression $^1/_{1+\exp(-(\beta_0 + \beta_1 \cdot x))}$. Note that since x has a finite number of values, p will too.

Smooth calibration for the scoring system

Take a random sample of n subjects $\{1,2,\ldots,i,\ldots,n\}$ where for each subject i we have x_i (the value of the score on a scoring system as defined above) and z_i (taking the value 1 if the subject has experienced an event and 0 otherwise). In turn, since we are using a scoring system, we have (using the above notation) $x_{min}, x_{max}, \beta_0$ and β_1, and in consequence p_i (probability of event).

For each subject i we now define the random variable $L_i = \beta_0 + \beta_1 \cdot x_i$. Smooth calibration consists of fitting a logistic regression model to the set $\{(z_i, L_i), i = 1, \ldots, n\}$ with the parameterization $logit(z) = a + f(L)$, where f is a smooth function of L, like splines or loess transformations, and a is the intercept of the model [5]. Through this new model we obtain the observed probabilities of the event and compare them with those predicted by the scoring system through a Cartesian graph. This graph will be represented in the space [0,1] x [0,1] and the straight line joining the points (0,0) and (1,1) will be added, as it represents the observed probabilities corresponding to those predicted by the scoring system. The smooth curve will be represented together with its associated confidence intervals, which can be obtained through bootstrapping [5]. Note that our system will have a total of $x_{max} - x_{min} + 1$ points represented on the Cartesian graph.

The estimated calibration index

The estimated calibration index (ECI) is a measure that has been proposed to determine the lack of calibration of a predictive model [5,12]. The ECI consists of calculating the mean

squared difference between the observed risk (obtained by smooth curves) and the risk predicted by the model in a total of N observations (by bootstrapping it would be in each of the samples). The ECI has a range of values from 0 to 100, where the null value corresponds to absolute perfection between the model and reality [5]. Although the ECI summarizes the lack of calibration in a single number, it has been observed that small values thereof (ECI = 1.67) produce models that are not well calibrated [12]. In other words, if a model is well calibrated, it will have a low ECI value, but the opposite does not hold true. In short, it is a necessary but not sufficient condition. Consequently, we have to represent the Cartesian graph of the models that obtain a low ECI.

The proposed algorithm to calculate the sample size to externally validate a scoring system

Using all the concepts defined above (AUC, scoring systems, smooth calibration and the ECI) we now detail an algorithm to evaluate how to calculate the sample size to externally validate a scoring system based on a logistic regression model, since we may have a sample size with a number of events and non-events different than 100, as is stated in the literature [5,8]. We must bear in mind that two aspects must be assessed (discrimination and calibration), the first of which does not need an excessive sample size to find statistically significant differences [8]. However, obtaining the sample size to determine if a model is well calibrated requires further study, using simulated samples [8]. For this reason, we will focus on the sample size for smooth calibration.

First, a few considerations; as noted above, the ECI is a measure that can help us with this task, since values close to 0 are necessary, but not sufficient, to say that a model is well calibrated. Therefore, we establish cut-off points near the null value for the ECI and determine if the model is well calibrated through the interpretation of the smooth calibration plot. We must also bear in mind that the random variable of the scores (x) can be considered to have a multinomial distribution since it has a finite number of values. In addition, we must have the proportion of subjects with an event (p_{event}), and then establish a possible range of values for the number of events (n_{event}) in order to check its calibration. With these considerations and the concepts discussed above, we can now detail the proposed algorithm:

1. Establish n_{event} (if it is the first time this step is initiated, n_{event} takes the minimum value of the possible range of values to check):
   a. Simulate a random sample from the vector (x,z) through the multinomial distribution of the scores and from the logistic regression model associated with the scoring system, with n_{event} subjects with the event and with $n_{non-event} = \frac{1-p_{event}}{p_{event}} \cdot n_{event}$ subjects without the event. Note that $n_{non-event}$ could have decimals, so we round it to the nearest whole number.
   b. In the sample in step 1a determine the AUC and observed event probabilities for each score through smooth curves.
   c. Repeat steps 1a and 1b a predetermined number of times N (for example, 1000 times) in order to construct the distribution of these parameters. Once the above steps have been repeated N times, continue with step 2.
2. Determine the ECI value with the total N observations performed, save the smooth calibration plot with the confidence intervals only and calculate confidence intervals for the AUC. Note that we are only interested in the confidence intervals in order to have a threshold for

possible population values, that is, with a high probability of ensuring that the model is well calibrated and can properly discriminate the subject with an event.

3. Recalculate n_{event} as $n_{event} = n_{event} + 1$ and go to step 1, unless we have already verified the full range of possible values for n_{event}, in which case we go to step 4.
4. With the cut-off points determined a priori for the ECI, create indicator variables to determine whether the number of events n_{event} verifies that the ECI is lower than these cut-off points.
5. Construct the ROC curves with n_{event} (quantitative variable) and the indicator variables in the ECI smaller than the cut-off points.
6. Determine the optimum point of n_{event} for each of the ECI cut-off points, as explained in the section on ROC curves.
7. Interpret the smooth calibration plots of the sample sizes obtained in step 6.
8. Set the sample size as the minimum value of n_{event} from step 6 that is properly calibrated.

Case study

We then applied the proposed algorithm to a scoring system for predicting mortality in the ICU [10]. The minimum score of this system is $x_{min} = 0$ points and the maximum score is $x_{max} = 15$ points, the coefficients of the logistic regression model associated with the system are $\beta_0 = -5.92252114678228$ and $\beta_1 = 0.6$, the proportion of events is $p_{event} = 0.1078199052132701421800947867298$6 and the probability distribution for each of the associated scores (ordered from $x_{min} = 0$ to $x_{max} = 15$) is (0.29023508137432200,0.03887884267631100,0.09222423146473780, 0.18625678119349000,0.05967450271247740,0.08318264014466550, 0.06057866184448460,0.02893309222423150,0.02441229656419530, 0.02622061482820980,0.03345388788426760,0.00994575045207957, 0.03526220614828210,0.02893309222423150,0,0.00180831826401447). These data were obtained from the original publication [10]. The established range of possible values for n_{event} was between 25 and 1000. The smooth curves were performed using linear splines.

To visualize the influence of sample size on discrimination and calibration, line graphs for the confidence intervals for the AUC and ECI were created. This evolution was analyzed using a video for soft calibration plots, which shows the adjustment to the perfect line of the curves with increasing n_{event}. The cut-off points established for the ECI were 2, 1.75, 1.5, 1.25, 1, 0.75, 0.5 and 0.25.

Results

Fig 1 shows the evolution of the AUC as the number of events in the sample increases, while Fig 2 represents the same evolution for the ECI. This evolution for smooth curves can be viewed in S1 Video. As can be seen, by increasing the sample size the errors are reduced and the bars of the smooth curve approach the perfect condition. These charts and the video indicate the presence of a certain point (number of patients) where we have a reduced error to carry out our external validation.

With the cut-off points chosen for the ECI (2, 1.75, 1.5, 1.25, 1, 0.75, 0.5 and 0.25), the number of events for the sample following our algorithm was 42, 51, 55, 56, 69, 167, 196 and 430, respectively. The smooth calibration plots for these sample sizes, along with the initial value of the verified range ($n_{event} = 25$), the value suggested in the literature ($n_{event} = 100$) and the final

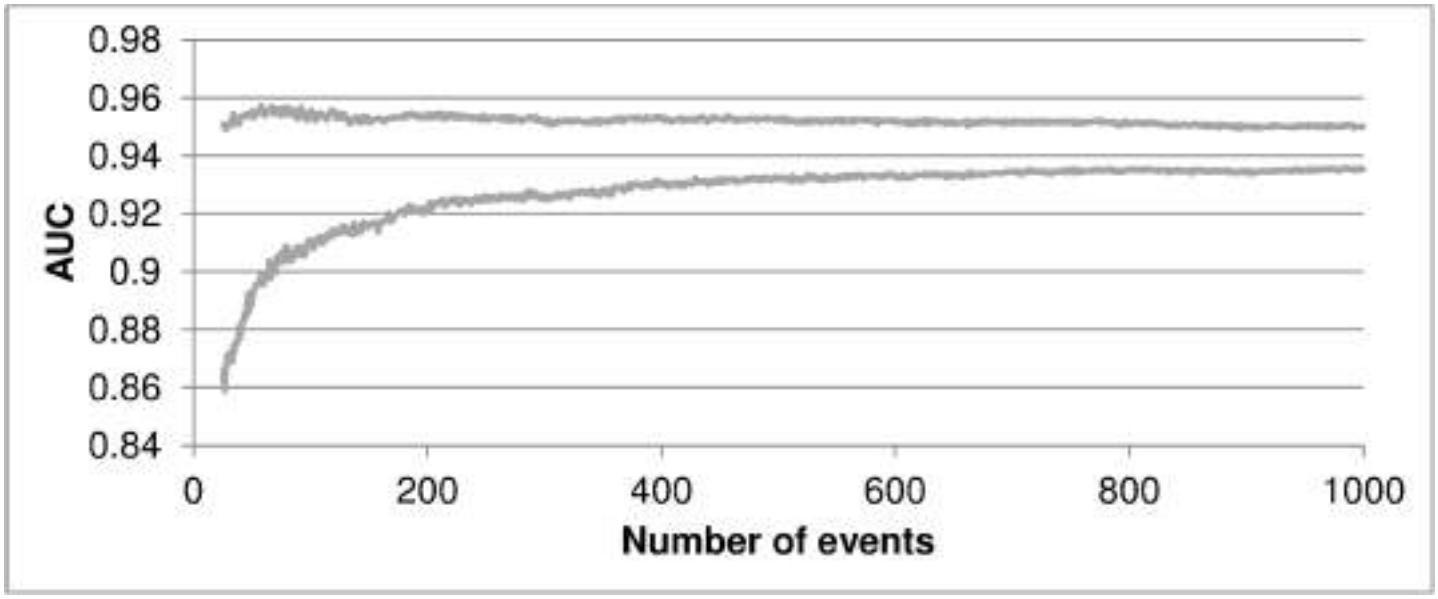


Fig 1. Confidence intervals for the area under the ROC curve according to the number of events in the sample. AUC, area under the ROC curve.

https://doi.org/10.1371/journal.pone.0176726.g001

value of the range (n_{event} = 1000), are shown in Fig 3. Note that these images are screenshots from the previous video with sample sizes predetermined by the algorithm; as the sample size increases the bars for the confidence intervals become closer to the perfect condition. Here we see that the minimum number of events that obtain good calibration is 69. This is complemented by an ECI<1.25. If we calculate the total number of patients in the sample through p_{event}, this is 640 patients (69 deceased and 571 living). This sample size would have 100 deceased and 828 living patients (938 in total) as recommended by the literature, representing 298 patients more than by following our algorithm.

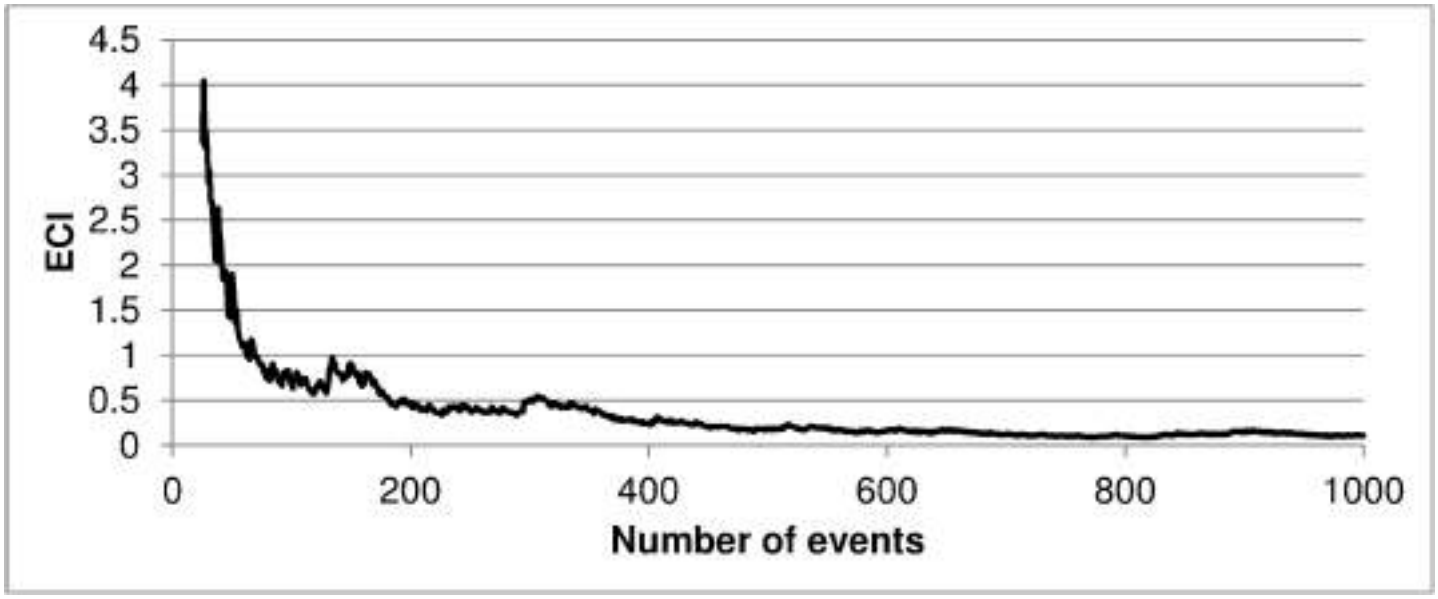


Fig 2. Estimated calibration index values according to the number of events in the sample. ECI, estimated calibration index.

https://doi.org/10.1371/journal.pone.0176726.g002

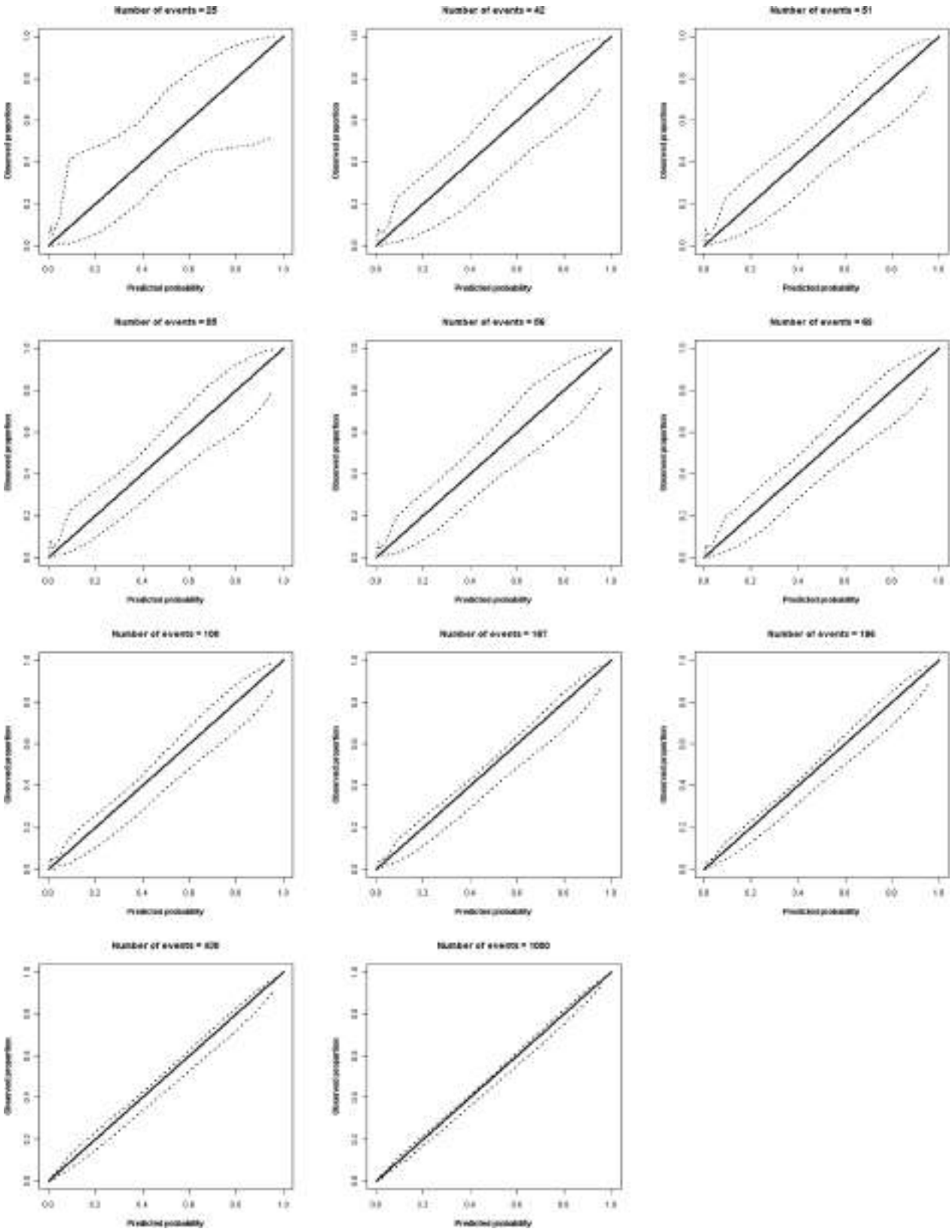


Fig 3. Smooth calibration plots (linear splines) for several sample sizes. The dashed lines denote the confidence intervals. The central line denotes the perfect prediction.

https://doi.org/10.1371/journal.pone.0176726.g003

Discussion

Our study developed an algorithm to calculate the sample size to externally validate a scoring system based on a binary logistic regression model, analyzing the smooth calibration plot and lack of calibration of this plot from the calculation of the ECI. As an example, the algorithm was applied to a scoring system to predict mortality in the ICU.

When comparing our algorithm with that published in the scientific literature, we note that other studies have considered a universal point (number of events/non-events = 100) [5,8]. As mentioned above, this is not entirely correct, because according to the predictive model that we are externally validating, the sample size for this validation should be independent [9]. However, we must bear in mind that the algorithm we have developed is for scoring systems, which are a specific case of binary logistic regression models.

Regarding the cut-off points of the ECI, we established this system to determine our sample size. However, another approach to this problem could be to consider all the calibration graphs and visually choose the graph that indicates the scoring system is properly calibrated. We wanted to incorporate the ECI because it is an objective way to measure lack of calibration and, when supplemented by the calibration graph, enabled us to view the issue in a more rigorous manner [5].

We recommend the use of our algorithm to calculate the sample size to externally validate scoring systems based on binary logistic regression models. Its application provides the number of patients required for the study, which may be fewer (or more) than 100 events and 100 non-events, as has been specified in the scientific literature [5,8].

According to the results of the case study performed (mortality in ICU) a sample size of 640 patients was obtained, which included 69 deaths. Consequently, if others plan to conduct studies to externally validate the scoring system to predict mortality in the ICU [10], the sample size calculation for these studies is available to them.

The main strength of this work is the algorithm developed to calculate the sample size to externally validate scoring systems based on binary logistic regression models. This subject has not been addressed in depth in the scientific literature, with the use of 100 events and 100 non-events being the recommendation, regardless of the characteristics of the model [5,8]. The value of 100 should not be fixed, however, because there are factors that have been shown to influence the calibration graph [9]. We also highlight the use of smooth curves rather than risk categorizations such as the Hosmer-Lemeshow test, as they give greater validity to the results [5]. Finally, we believe that this algorithm can be extended to more complex cases, such as scoring systems based on survival models or logistic regression models/overall survival, since scoring systems are a specific case of the same.

As a limitation, we note that this calculation carries a high computational cost due to the necessity of multiple bootstrapping samples for each number of events from the proposed range. In our case, our range had 976 possible values and 1000 simulations in each value, equivalent to a total of 976,000 simulations, in which the AUC and the observed values were calculated through smooth curves. However, if we consider the benefit that the use of this algorithm can provide, this would not be a limitation. In our example we have reduced the sample size suggested by the literature by 298 patients, which corresponds to a substantial reduction in both economic costs and the time needed to recruit study participants. In other words, the algorithm is useful to assess the issue being studied (sample size to validate scoring systems based on binary logistic regression models).

As a new line of research, we propose adapting this algorithm to a scoring system based on survival models. To do this, we will need to set cut-off points for prediction time and obtain the observed event probabilities of these cut-off points through smooth curves. These

probabilities will depend on the corresponding value in the scoring system and the baseline survival at the time being assessed [2]. We encourage other authors to adapt our algorithm to general logistic regression models.

Conclusions

This paper provides an algorithm to determine the sample size for validating scoring systems based on binary logistic regression models. The algorithm is based on bootstrapping and basic concepts when validating a predictive model (ROC curve, smooth calibration plots, and ECI). We applied the algorithm to a case to help readers better understand its application.

Supporting information

S1 Video. Smooth calibration plots for the example (number of events from 25 to 1000).
(MP4)

Acknowledgments

The authors thank Maria Repice and Ian Johnstone for their linguistic collaboration in the English version of the final text.

Author Contributions

Conceptualization: AP DMF EC MTL VFG.

Data curation: AP MTL.

Formal analysis: AP.

Investigation: AP DMF VFG.

Methodology: AP DMF VFG.

Project administration: AP.

Resources: VFG.

Software: AP DMF.

Supervision: AP.

Validation: AP DMF.

Visualization: AP DMF EC MTL VFG.

Writing – original draft: AP.

Writing – review & editing: AP DMF EC MTL VFG.

References

1. Hosmer DW, Lemeshow S. Applied logistic regression. New York, USA: Wiley; 2000.
2. Sullivan LM, Massaro JM, D'Agostino RB Sr. Presentation of multivariate data for clinical use: the Framingham study risk score functions. Stat Med. 2004; 23: 1631–1660. https://doi.org/10.1002/sim.1742 PMID: 15122742
3. Steyerberg EW. Clinical prediction models. A practical approach to development, validation, and updating. New York, USA: Springer-Verlag; 2009.

4. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143: 29–36. https://doi.org/10.1148/radiology.143.1.7063747 PMID: 7063747
5. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016; 74: 167–176. https://doi.org/10.1016/j.jclinepi.2015.12.005 PMID: 26772608
6. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015; 162: 55–63. Erratum in: Ann Intern Med. 2015; 162: 600. https://doi.org/10.7326/M14-0697 PMID: 25560714
7. Chow S, Wang H, Shao J. Sample Size Calculations in Clinical Research. 2nd ed. New York, USA: Chapman & Hall/CRC; 2008.
8. 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: 214–226. https://doi.org/10.1002/sim.6787 PMID: 26553135
9. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med. 2014; 33: 517–535. https://doi.org/10.1002/sim.5941 PMID: 24002997
10. Dólera-Moreno C, Palazón-Bru A, Colomina-Climent F, Gil-Guillén VF. Construction and internal validation of a new mortality risk score for patients admitted to the intensive care unit. Int J Clin Pract. 2016.
11. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978; 8: 283–298. PMID: 112681
12. Van Hoorde K, Van Huffel S, Timmerman D, Bourne T, Van Calster B. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. J Biomed Inform. 2015; 54: 283–293. https://doi.org/10.1016/j.jbi.2014.12.016 PMID: 25579635

TITLE PAGE

Title: A method to validate scoring systems based on logistic regression models to predict binary outcomes via a mobile application for Android with an example of a real case.

Authors: David Manuel Folgado-de la Rosa[a], Antonio Palazón-Bru[b], Vicente Francisco Gil-Guillén[b].

Institutions:

a. Department of Data Science/Scoring, CrossLend GmbH, Berlin, Berlin, Germany.

b. Department of Clinical Medicine, Miguel Hernández University, San Juan de Alicante, Alicante, Spain.

Corresponding author: Prof. Antonio Palazón-Bru, PhD. Department of Clinical Medicine, Miguel Hernández University, Carretera de Valencia - Alicante S/N, 03550 San Juan de Alicante, Spain. Telephone: +34 965919449. Fax: +34 965919450. E-mail: antonio.pb23@gmail.com

Word count: 3136.

ABSTRACT

Background and objectives: To use a points system based on a logistic regression model to predict a binary event in a given population, the validation of this system is necessary. The most correct way to do this is to calculate discrimination and calibration using bootstrapping. Discrimination can be addressed through the area under the receiver operating characteristic curve (AUC) and calibration through the representation of the smoothed calibration plot (most recommended method). As this is not a simple task, we developed a methodology to construct a mobile application in Android to perform this task.

Methods: The construction of the application is based on source code written in language supported by Android. It is designed to use a database of subjects to be analyzed and to be able to apply statistical methods widely used in the scientific literature to validate a points system (bootstrap, AUC, logistic regression models and smooth curves). As an example our methodology was applied on simulated points system data (doi: 10.1111/ijcp.12851) to predict mortality on admission to intensive care units (Google Play: *ICU mortality*). The results were compared with those obtained applying the same methods in the R statistical package.

Results: No differences were found between the results obtained in the mobile application and those from the R statistical package, an expected result when applying the same mathematical techniques.

Conclusions: Our methodology may be applied to other point systems for predicting binary events, as well as to other types of predictive models.

KEYWORDS: Mobile Applications; Models, Statistical Software; Validation; Validation Studies as Topic.

1. INTRODUCTION[1]

In biomedical research, one of the most commonly used mathematical models to calculate the probability of a binary event (outcome), is the binary logistic regression model.[1] This model attempts to predict this event (dependent variable) based on a set of independent variables, integrating arithmetic operations and exponential functions.[1] Thus, it is difficult to calculate the probability of our outcome without using an electronic device. Bearing in mind that in clinical practice decisions must be made immediately, the researchers of the Framingham Heart Study developed an algorithm to transform the binary logistic regression model into a points system that can be integrated into routine clinical practice.[2]

The algorithm developed in the Framingham Heart Study associates a score (integer) to the independent variables, through a weighting of the estimated coefficients of the model, and a total score for each subject is determined through the sum of all the scores associated with each of the independent variables of the model. This total score will have an associated probability of occurrence of the outcome in question, which is very similar to that obtained by the logistic regression model formula.[2]

Once the points system has been constructed in a given population, its external validation is necessary for use in other populations.[3] To do this, we must check discrimination and calibration.[4] The former is generally addressed by calculating the area under the receiver operating characteristic (ROC) curve (AUC), [5,6] and the latter has traditionally been done by calculating observed and expected events in risk groups.[1,7-9] We can also assess this issue using a logistic recalibration framework, which consists of fitting the following logistic regression model: $\text{logit}(Z) = a + b_L \cdot L$, which is used to estimate the observed event

[1] Area under the receiver operating characteristic curve (AUC); intensive care unit (ICU); receiver operating characteristic (ROC).

probabilities. If $b_L = 1$ (calibration slope) and $a = 0$, the model coincides with the diagonal line; that is, the observed event probabilities coincide with the expected probabilities (calibration).[4] Finally, a flexible calibration curve can be fit to estimate these observed risks, $\text{logit}(Z) = a + f(L)$, where $f(L)$ is a continuous function in the L parameter, such as spline or loess transformations.[4] Of all these calibration techniques, smooth curves are the most recommended.[4]

Traditionally, validation studies of points systems for a binary event have been performed on a single sample of patients;[10-12] i.e., the AUC and the calibration have been calculated on that single validation sample. However, currently, the computational capacity of computers has evolved and this has led to the validation of points systems for a binary event through bootstrap samples.[9,13-16] The enormous computing power of modern computers considerably facilitates the applicability of this computationally expensive method.

Bearing in mind that the use of a points system of these characteristics in a given population requires validation and that this is not a simple task, a scheme is proposed for the development of an application for mobile phones using the Android system in which health professionals can validate the points system through bootstrap samples and, once validated, can apply the system to their patients. To do this, the clinician enters patient data until reaching the sample size necessary to externally validate a predictive model.[17,18] Once this number of subjects has been reached, the mobile application will display the results of the validation. After this validation, the clinician will be able to apply the points system to a new patient in his or her usual clinical practice.

To clarify to the readers the development of how to construct the mobile application, the methodology of the article will be structured in four differentiated parts. In the first part, the statistical methods to be applied will be indicated, describing their vector form, since it is a

key point when programming the calculations in the application. The second part will explain the scheme of operation of the application through flow diagrams that help the reader to understand it. Thirdly, this scheme will be implemented in an already published points system, which predicts mortality in intensive care units (ICUs).[19] Finally, in order for readers to verify that there are no differences between performing the validation with statistical packages or the mobile application, the results between the two methods will be compared in a sample of simulated data (mobile application and R statistical package).

2. METHODS

2.1. A review of the statistical methods to be included in the mobile application

2.1.1. Definition of a scoring system

Each subject i ($i = 1,...,n$) has a score of $x_i \in \{\tilde{x},..., -1, 0, 1,...,x^*\}$, with x^* being the highest possible score and $\tilde{x}$ being the lowest. At the same time x_i has an associated probability of event p_i, such that $p_i = (1 + \exp(-(\beta_0 + \beta_1 \cdot x_i)))^{-1}$, where β_0 and β_1 are the regression coefficients associated with the points system.[2] We define z_i as the event indicator variable, that is, it takes the value of 1 when the subject i has an event and 0 otherwise.

2.1.2. Area under the ROC curve for a scoring system

Based on the above notation, we define $E = \{i : z_i = 1, i = 1,...,n\}$ and $\overline{E} = \{i : z_i = 0, i = 1,...,n\}$, such that $E \cup \overline{E} = \{1,...,n\}$ and $E \cap \overline{E} = \emptyset$. The ROC curve is defined as the union of the following points on a Cartesian graph [5]:

$$\left(1 - \frac{|i \in \overline{E} : x_i < x|}{|\overline{E}|}, \frac{|i \in E : x_i \geq x|}{|E|}\right) x \in \{\tilde{x},..., -1, 0, 1,...,x^* + 1\},$$

where $|\cdot|$ is the cardinal function of a given set, i.e. the number of elements of that set.

Now we calculate the area under this curve $[0,1]\ x\ [0,1]$ (AUC). Let $j \in E$ and $l \in \overline{E}$ be defined as:

$$S(j,l) = \begin{cases} 1 & if\ x_j > x_l \\ 1/2 & if\ x_j = x_l. \\ 0 & if\ x_j < x_l \end{cases}$$

The AUC calculation, given that the total score is a discrete variable, is obtained with the following expression [5]:

$$AUC = \frac{1}{|E| \cdot |\overline{E}|} \cdot \sum_{j \in E} \sum_{l \in \overline{E}} S(j,l).$$

When interpreting the value of the AUC, the closer this is to one indicates that the points system discriminates to a greater extent between which subject experiences an event and which subject does not.

2.1.3. Smooth calibration

Using the notation above, for each subject i we define $L_i = \beta_0 + \beta_1 \cdot x_i$. Next, we fit a binary logistic regression model using as an explanatory variable a flexible curve dependent on the value of L. This can be constructed using differentiable curves defined piecewise, like spline or loess transformations.[4] Through this model we estimate the observed event probabilities for each possible score of the points system. These observed probabilities will be compared with those predicted by our system through a Cartesian graph, which will be used to determine whether the probability curve (union of all points) fits the diagonal line (calibration).

2.1.4. Bootstrapping methodology and its application to scoring systems

Given a sample of n subjects, we define the bootstrap sample as a random sample with replacement of a total of n subjects obtained from the original sample. This random sample can be taken using various designs, such as simple random sampling or stratified sampling. In other words, there will be elements repeated more than once and other elements that will not belong to the sample.

Bootstrapping methodology consists of obtaining a large number of bootstrap samples, usually 1000 samples, and determining in each of them the value of a statistical parameter. Through these values we construct the distribution of that parameter.[15] If we apply this methodology when validating a points system, we will be able to construct the distribution of the AUC, in addition to obtaining the smooth calibration plot. The AUC indicates the discriminatory capacity of the points system to differentiate between subjects who experience an event and those who do not. Thus, if the values of its distribution are close to one, this means that the discrimination of our points system is satisfactory. In addition, we can say that the points system has been calibrated if the smooth calibration plot conforms to the diagonal line (observed risks are similar to those predicted by the points system). In short, if both discrimination and validation are satisfactory, we can say that our points system has been externally validated in our population.

To study the distribution of the AUC we carry out a descriptive analysis to determine its mean and standard deviation. Following the analysis of these values and of the smooth calibration plot, we can conclude whether or not we can use our points system in our population, i.e. whether the points system has been validated.

2.2. Construction of the mobile application for Android

The source code is written in a programming language, Java or Kotlin, in the Android Studio integrated development environment, where it is structured in four main components: 1)

Setup, 2) *Database*, 3) *Validation* and 4) *Predictor*. First, *Setup* indicates the initial situation of the points system in the study population, indicating whether the system is validated and defining the sample size. The introduction of patient data is handled in the following component (*Database*). In *Validation*, we apply the statistical methods to perform the external validation of the points system. Finally, if the system has been validated, either by the investigator in a previous study or through the application, *Predictor* applies it to a new patient determining the risk of the binary event of interest.

Now that we have detailed the structure, we will explain the mechanism of operation of the mobile application through flow diagrams (Figs. 1-3). The first thing we must do is to tell the application whether the system is validated (Fig. 1). If so, the application will allow us to calculate the score and the probability of an event in a patient. Conversely, if the system is not validated, we must calculate the sample size, which should include at least 100 subjects with an event and a number of subjects without an event proportional to that of the population. This means that if in our population this proportion is $\tilde{p}$, we should have at least $\left[\frac{1-\tilde{p}}{\tilde{p}}\cdot 100\right]$ subjects without an event. However, the figure of 100 events is at a general level for any predictive model regardless of its characteristics, but since 2017 an algorithm has been available that determines the sample size to validate particular points systems,[18] indicating as an example that for the ICU points system with 69 events and $\left[\frac{1-\tilde{p}}{\tilde{p}}\cdot 69\right]$ patients without an event, a good sample size was available. In the app the user is given the option to choose either of the two approaches. Next, the user must enter patient data until this sample size is reached, after which the application will begin the validation process.

The user is offered an interface for the creation of the database, through which the user can delete, update and add patients (Fig. 2). The database, after calculating the score through the system variables, will have the following components: score (x), event indicator (z), event

probability (p) and linear predictor (L), which will be transformed into spline functions. Note that the value of n (total number of subjects entered thus far in the database) (Fig. 2) varies according to each of these operations. Once we have introduced the necessary sample size, the mobile application will proceed to the validation of the points system using bootstrap methodology (Fig. 3). To facilitate this task, it is recommended to establish an event and non-event patient counter, in addition to providing the list of patients entered up to that moment, to allow the user to edit or delete records.

To validate the points system (Fig. 3), we start from the database introduced with n elements, from a counter variable (i), from an empty vector with 100 components to store the values of the AUC distribution, and from an empty matrix with 100 rows and $x^* - \tilde{x} + 1$ columns to store the observed probabilities for each possible score in each bootstrap sample. This counter variable (i) performs a total of 100 iterations, obtaining a bootstrap sample in each of them in which it calculates the AUC. In other words, since it is performed 100 times, the distribution of this parameter is obtained and is used to determine whether the points system correctly discriminates the event in the analyzed population. As mentioned above, the AUC must have a value close to one to be able to say that it correctly discriminates which patient will experience an event. This will be assessed through the calculation of the mean and standard deviation. The application also determines in each bootstrap sample the observed event probabilities for each score of the system through smooth curves, with which the calibration plot is constructed. This plot must be evaluated by the user according to the fit of the curve to the diagonal line (observed=expected). A point to keep in mind is that smooth calibration may not converge in some points systems, so another calibration method would have to be applied, such as linear calibration. However, some guidelines should be indicated to allow interpretation of the results, according to the existing dispersion between the observed and the

expected outcome. This information, together with the database entered, should be sent to the user of the application by email or similar method.

In the event that we have obtained satisfactory results to conclude that the system has been externally validated in our population, just as if when starting the system we have indicated that the system was previously validated, the user is allowed to apply the system to a new patient, introducing the patient variables and determining the outcome risk.

2.3. System applied to a simulated data set (ICU mortality)

To make it easier for the reader to understand the algorithm for developing the mobile application, it has been applied to the points system for predicting mortality in the ICU.[19] This system assigns a total score to each subject, based on the sum of the partial scores of the variables: medical admission, sepsis, inotropic support, cardiology admission, mechanical ventilation, function scale (independent, dependent and disability).[19,20] The mobile application has been uploaded to Google Play under the name *ICU mortality* and its download is free for all users. The application performs smooth calibration through spline transformations unless there is no convergence, in which case the application will perform linear calibration. Fig. 4 shows screen captures of the different phases of the mobile application, such as the report of the validation.

With respect to using the application to validate the proposed points system, with the sole purpose of being able to visualize the entire process without entering any data, we have simulated a database similar to the one on which the points system was developed. This simulation is based on logistic regression models the coefficients of which were obtained with the original database.[19] For the reader to be able to replicate our data set, both the corresponding R code and the database generated by this code and used to show the results obtained (Supplementary File) have been incorporated as supplementary material. The

simulated sample (fixtures) can be loaded into the application by pressing seven times on the counter that shows the number of patients that have been entered. It will have the necessary sample size for validation, with the exception of a deceased patient, so that the user can enter it manually and see the complete process. Finally, it should be indicated that the application calculates the matrix of grade 1 splines (B-spline basis matrix), using as nodes the 20th, 40th, 60th and 80th percentiles.

2.4. Comparison of the results obtained using the statistical software and the mobile application

The distributions obtained using both methods (R and mobile application) for our parameters (AUC and smooth probabilities for each score) were compared descriptively and graphically.

3. RESULTS

When comparing the results obtained with the R statistical package and with the mobile application, almost no differences were found in either the AUC (R, 0.94±0.01; mobile application, 0.93±0.01) or the calibration (Fig. 5). Firstly, the AUC values differed by just one hundredth, which is irrelevant and could be due to the bootstrap samples selected or even to the rounding of the values themselves so as to have two significant figures. Second and last, the smooth curves of the two procedures clearly overlap, which shows that the values of the observed probabilities are very similar. All this was logical and expected, since the mathematical methods used were the same in these procedures (R and our mobile application).

4. DISCUSSION

In this article we describe an algorithm that provides readers with a set of guidelines to develop an application for mobile phones in the Android system and to validate any points system based on a logistic regression model to predict a binary event. This algorithm was applied on simulated data, and the results obtained using the R statistical package and using the mobile application were confirmed to show no differences, which was evident, as the same mathematical techniques were used.

The main strength of this study is the development of an algorithm that enables the automatic validation of points systems to predict binary events based on logistic regression models. Simply by using a phone with an Android operating system and executing our mobile application, these points systems may be implemented quickly in routine clinical practice. As a limitation, concerning the validation through bootstrapping, it is possible that the user may not be able to use his or her phone to its full capacity for other functions while the validation is being performed as this requires multiple mathematical calculations. Nonetheless, since mobile phone processors are becoming increasingly more powerful, this usage will be reduced rapidly with the implementation of new processors in mobile telephony. Furthermore, if we consider the time it may take to enter the characteristics of subjects in a database on a computer and then apply validation techniques through a statistical package, our mobile application usage minutes may account for just a small percentage of that time.

It could be said that the methodology to construct our mobile application needs validation. However, if we bear in mind that the mathematical techniques employed have been widely used in the scientific literature to validate points systems and that our algorithm adheres to these techniques, no such validation study is necessary.[3] In other words, our methodology can be applied to any points system to predict a binary event through a logistic regression model. This model is implemented in a mobile application, which allows any user to validate a points system, and if satisfactory results are obtained in this validation, the user can

implement the system in routine clinical practice. Finally, since the application is available to all users (Google Play: *ICU mortality*), we encourage other researchers to validate this points system in their populations, as well as to implement our methodology in other points systems.

In conclusion, in this article we have described a methodology to develop a mobile application for the Android operating system that validates points systems based on logistic regression models to predict binary events through bootstraping. This validation is based on mathematical techniques widely used in the scientific literature. To help the reader understand our methodology, a practical example is incorporated, and in this example, the validation results are obtained, both through the mobile application and through the R statistical package.

ACKNOWLEDGEMENTS

The authors thank Maria Repice and Ian Johnstone for their help with the English version of the final text.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

ETHICAL CONSIDERATIONS

We have used simulated data for this paper, consequently the ethical approval by an Institutional Review Board was not necessary.

REFERENCES

1. D.W. Hosmer, S. Lemeshow, Applied logistic regression (Wiley, New York NY, 2000).

2. L.M. Sullivan, J.M. Massaro, R.B. D'Agostino Sr, Presentation of multivariate data for clinical use: The Framingham Study risk score functions, Stat. Med. 23 (2004) 1631-1660.

3. A. Palazón-Bru, J.A. Carbayo-Herencia, M.I. Vigo, V.F. Gil-Guillén, A method to construct a points system to predict cardiovascular disease considering repeated measures of risk factors, PeerJ 4 (2016) e1673.

4. B. Van Calster, D. Nieboer, Y. Vergouwe, B. De Cock, M.J. Pencina, E.W. Steyerberg, A calibration hierarchy for risk models was defined: from utopia to empirical data, J. Clin. Epidemiol. 74 (2016) 167-176.

5. J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29-36.

6. D.M. Lloyd-Jones, Cardiovascular risk prediction: basic concepts, current status, and future directions, Circulation 121 (2010) 1768-1777.

7. R.B. D'Agostino Sr, R.S. Vasan, M.J. Pencina, P.A. Wolf, M. Cobain, J.M. Massaro, W.B. Kannel, General cardiovascular risk profile for use in primary care: the Framingham Heart Study, Circulation 117 (2008) 743-753.

8. M.J. Pencina, R.B. D'Agostino Sr, M.G. Larson, J.M. Massaro, R.S. Vasan, Predicting the 30-year risk of cardiovascular disease: the framingham heart study, Circulation 119 (2009) 3078-3084.

9. R.B. Schnabel, L.M. Sullivan, D. Levy, M.J. Pencina, J.M. Massaro, R.B. D'Agostino Sr, C. Newton-Cheh, J.F. Yamamoto, J.W. Magnani, T.M. Tadros, W.B. Kannel, T.J. Wang, P.T.

Ellinor, P.A. Wolf, R.S. Vasan, E.J. Benjamin, Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study, Lancet 373 (2009) 739-745.

10. T.J. Wang, J.M. Massaro, D. Levy, R.S. Vasan, P.A. Wolf, R.B. D'Agostino, M.G. Larson, W.B. Kannel, E.J. Benjamin, A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community: the Framingham Heart Study, JAMA 290 (2003) 1049-1056.

11. J. Marrugat, I. Subirana, E. Comín, C. Cabezas, J. Vila, R. Elosua, B.H. Nam, R. Ramos, J. Sala, P. Solanas, F. Cordón, J. Gené-Badia, R.B. D'Agostino; VERIFICA Investigators, Validity of an adaptation of the Framingham cardiovascular risk function: the VERIFICA Study. J. Epidemiol. Community. Health 61 (2007) 40-47.

12. P.W. Wilson, J.B. Meigs, L. Sullivan, C.S. Fox, D.M. Nathan, R.B. D'Agostino Sr, Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study, Arch. Intern. Med. 167 (2007) 1068-1074.

13. K. Chien, T. Cai, H. Hsu, T. Su, W. Chang, M. Chen, Y. Lee, F.B. Hu, A prediction model for type 2 diabetes risk among Chinese people, Diabetologia 52 (2009) 443-450.

14. L. Liu, Z. Tang, X. Li, Y. Luo, J. Guo, H. Li, X. Liu, L. Tao, A. Yan, X. Guo, A Novel Risk Score to the Prediction of 10-year Risk for Coronary Artery Disease Among the Elderly in Beijing Based on Competing Risk Model, Medicine (Baltimore) 95 (2016) e2997.

15. B. Efron, Bootstrap methods: Another look at the jackknife, Ann. Stat. 7 (1979) 1-26.

16. E.W. Steyerberg, F.E. Harrell Jr, G.J. Borsboom, M.J. Eijkemans, Y. Vergouwe, J.D. Habbema, Internal validation of predictive models: efficiency of some procedures for logistic regression analysis, J. Clin. Epidemiol. 54 (2001) 774-781.

17. G.S. Collins, E.O. Ogundimu, D.G. Altman, Sample size considerations for the external validation of a multivariable prognostic model: a resampling study, Stat. Med. 35 (2016) 214-226.

18. A. Palazón-Bru, D.M. Folgado-de la Rosa, E. Cortés-Castell, M.T. López-Cascales, V.F. Gil-Guillén, Sample size calculation to externally validate scoring systems based on logistic regression models, PLoS One 12 (2017) e0176726.

19. C. Dólera-Moreno, A. Palazón-Bru, F. Colomina-Climent, V.F. Gil-Guillén, Construction and internal validation of a new mortality risk score for patients admitted to the intensive care unit. Int. J. Clin. Pract. 70 (2016) 916-922.

20. K. Rockwood, X. Song, C. MacKnight, H. Bergman, D.B. Hogan, I. McDowell, A. Mitnitski, A global clinical measure of fitness and frailty in elderly people, CMAJ 173 (2005) 489-495.

FIGURE LEGENDS

Figure 1: General scheme of the mobile application.

Figure 2: Tasks to create the database in the mobile application.

n, number of patients in the database.

Figure 3: Tasks to validate the scoring system to predict binary outcomes in the mobile application.

AUC, area under the ROC curve; i, counter variable; sd, standard deviation.

Figure 4: Screenshots of the mobile application in its different stages.

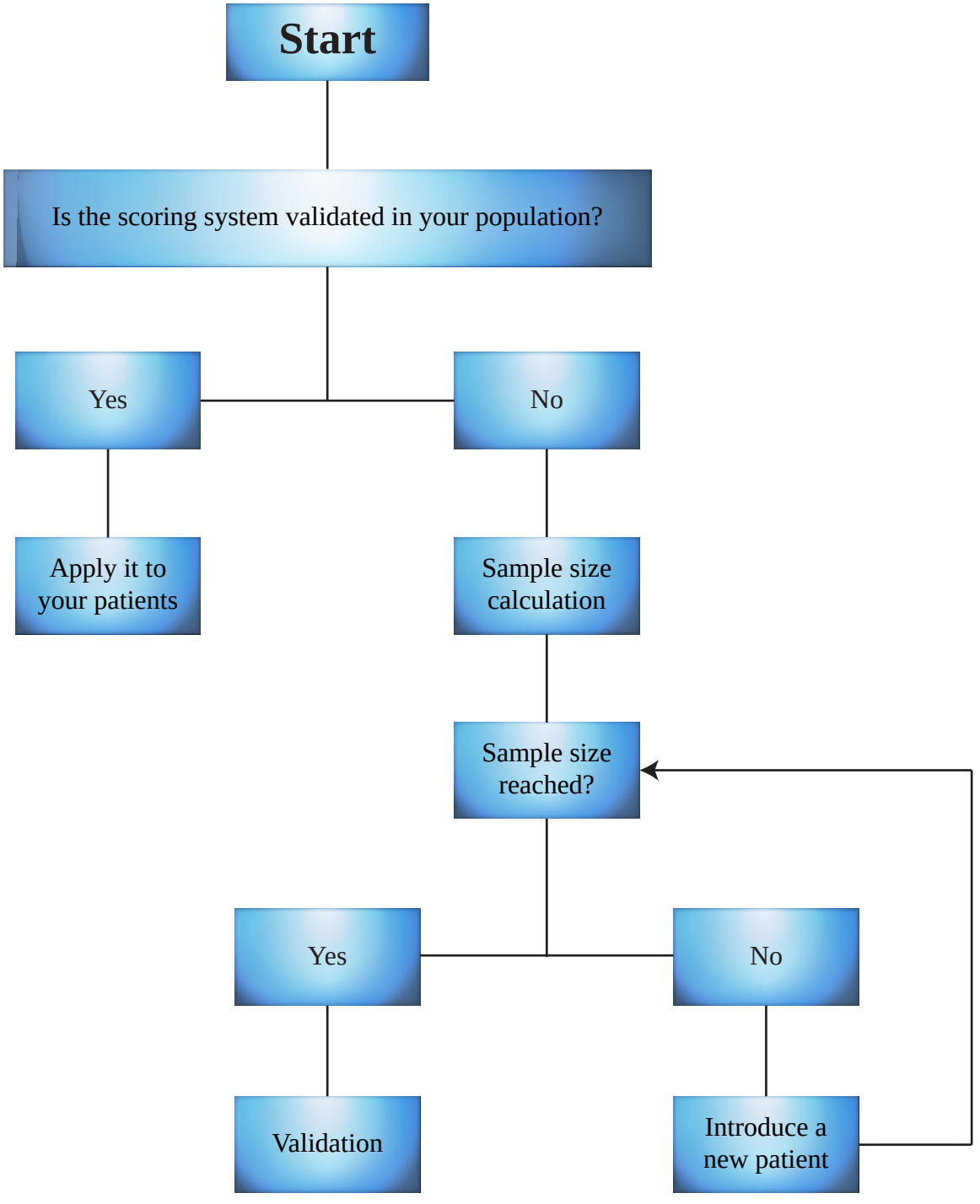
A, sample size calculation; B, creation of the database; C, report of the validation; D, report of the risk score after its validation.

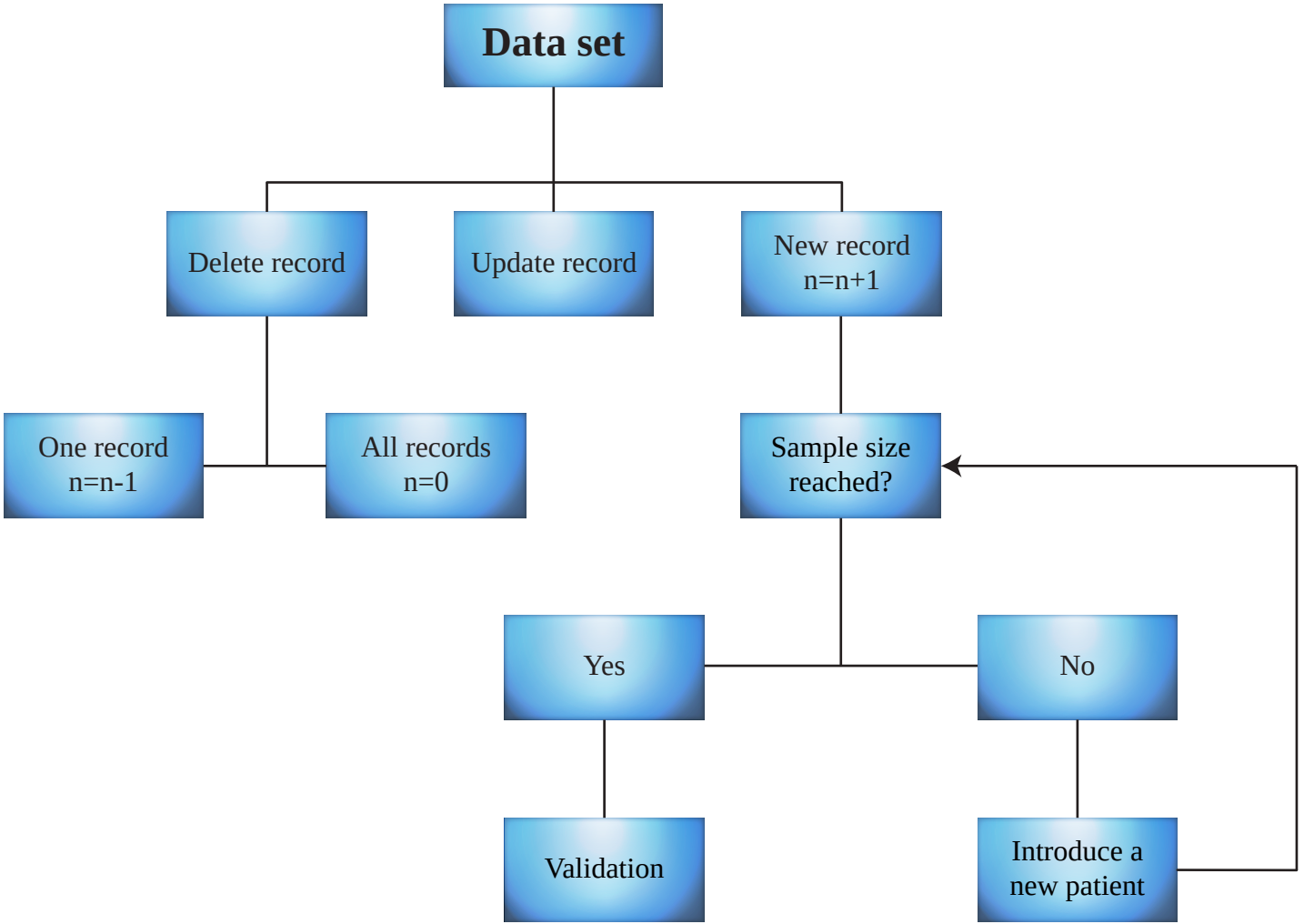
Figure 5: Smooth calibration plot using linear splines.

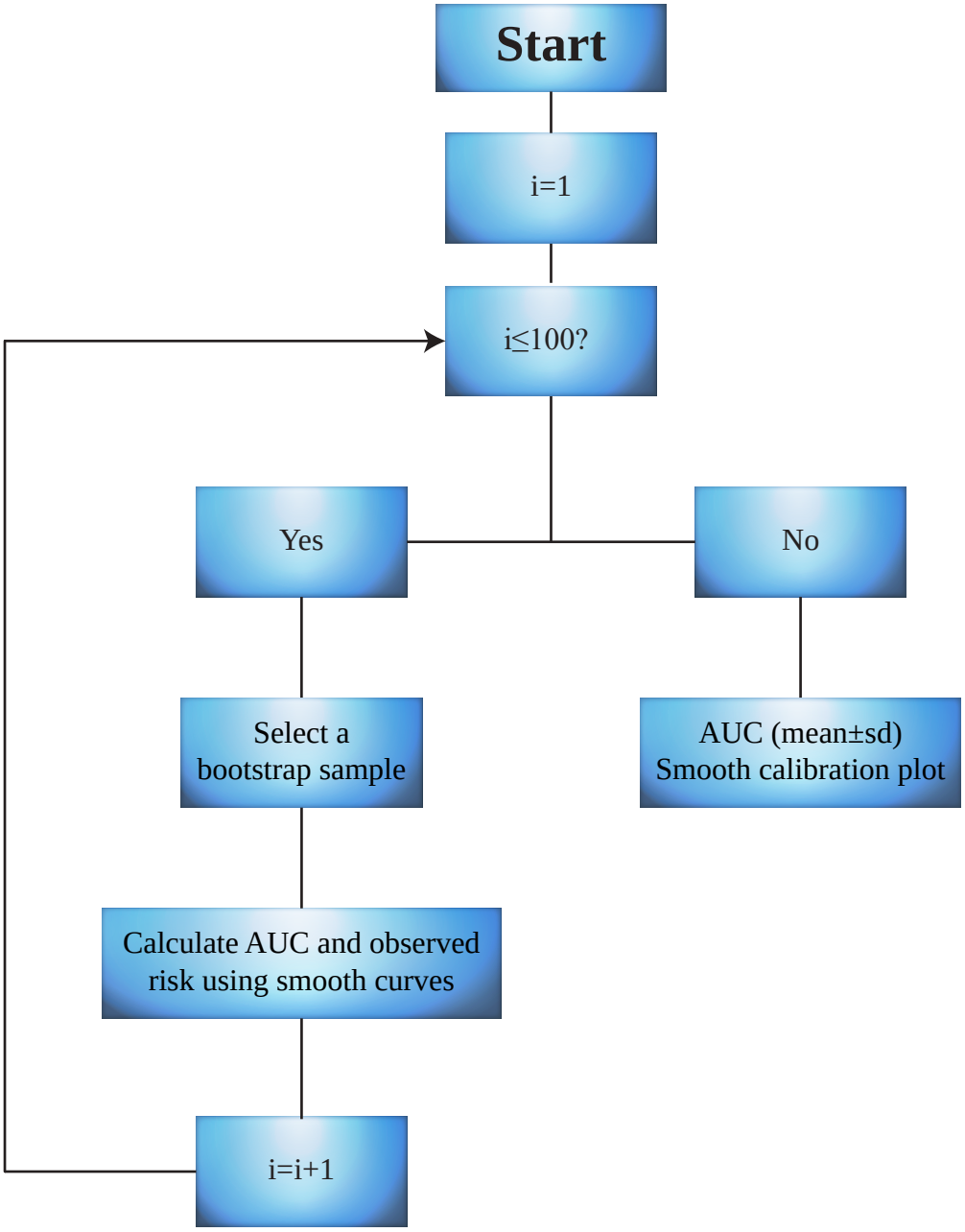
The thinner black line denotes the perfect prediction. The blue line is the model calibration using linear splines for the statistical software. The orange line denotes the same for the mobile application.

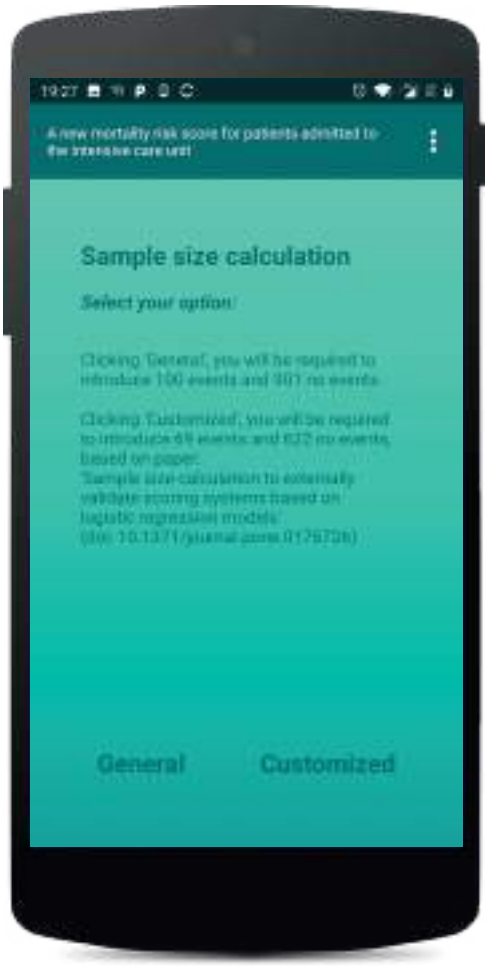
SUPPLEMENTAL MATERIAL

Simulated data set, and its R script, which was used to obtain the results provided in this paper.

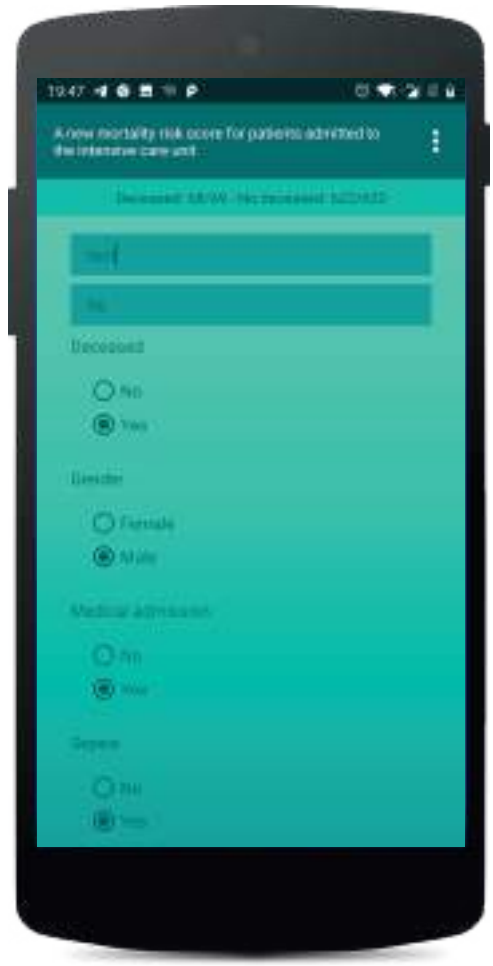

Start
Is the scoring system validated in your population?
Yes
No
Apply it to your patients
Sample size calculation
Sample size reached?
Yes
No
Validation
Introduce a new patient


Data set
Delete record
Update record
New record
n=n+1
One record
n=n-1
All records
n=0
Sample size
reached?
Yes
No
Validation
Introduce a
new patient

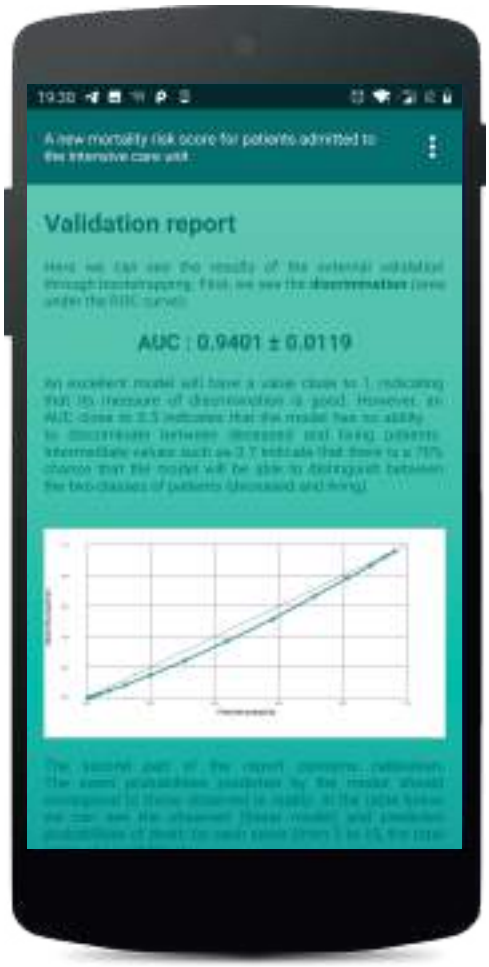

Start
i=1
i≤100?
Yes
No
Select a bootstrap sample
AUC (mean±sd) Smooth calibration plot
Calculate AUC and observed risk using smooth curves
i=i+1


A new mortality risk score for patients admitted to the intensive care unit
Sample size calculation
Select your option:
Clicking 'General', you will be required to introduce 100 events and 901 no events.
Clicking 'Customized', you will be required to introduce 69 events and 622 no events, based on paper:
'Sample size calculation to externally validate scoring systems based on logistic regression models'
(doi: 10.1371/journal.pone.0176726)
General
Customized

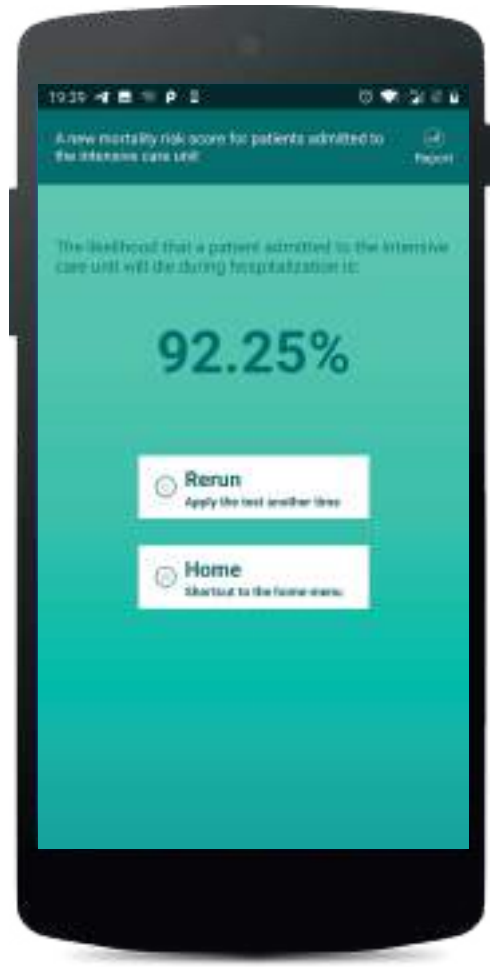
A


A new mortality risk score for patients admitted to the intensive care unit
Deceased
No
Yes
Gender
Female
Male
Medical admission
No
Yes
Sepsis
No
Yes

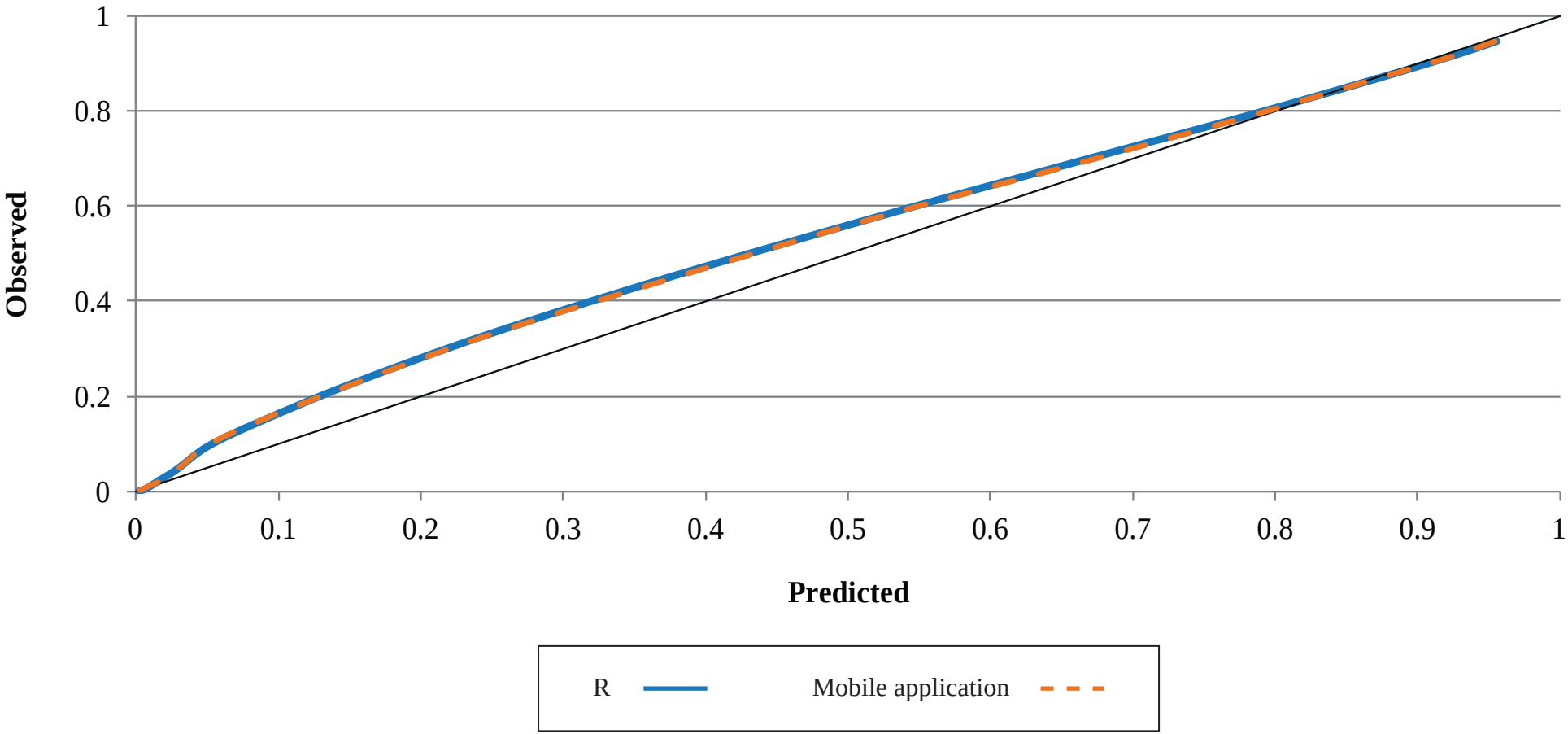
B


A new mortality risk score for patients admitted to the intensive care unit
Validation report
AUC : 0.9401 ± 0.0119

C


A new mortality risk score for patients admitted to the intensive care unit
Report
92.25%
Rerun
Home

D


Observed
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Predicted
R
Mobile application

The authors declare no conflict of interest.