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

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.

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.

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: $E = \{i: z_i = 1, i = 1,\ldots,n\}$ 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},β_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 $p_{event} = 0.10781990521327014218009478672986$ 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

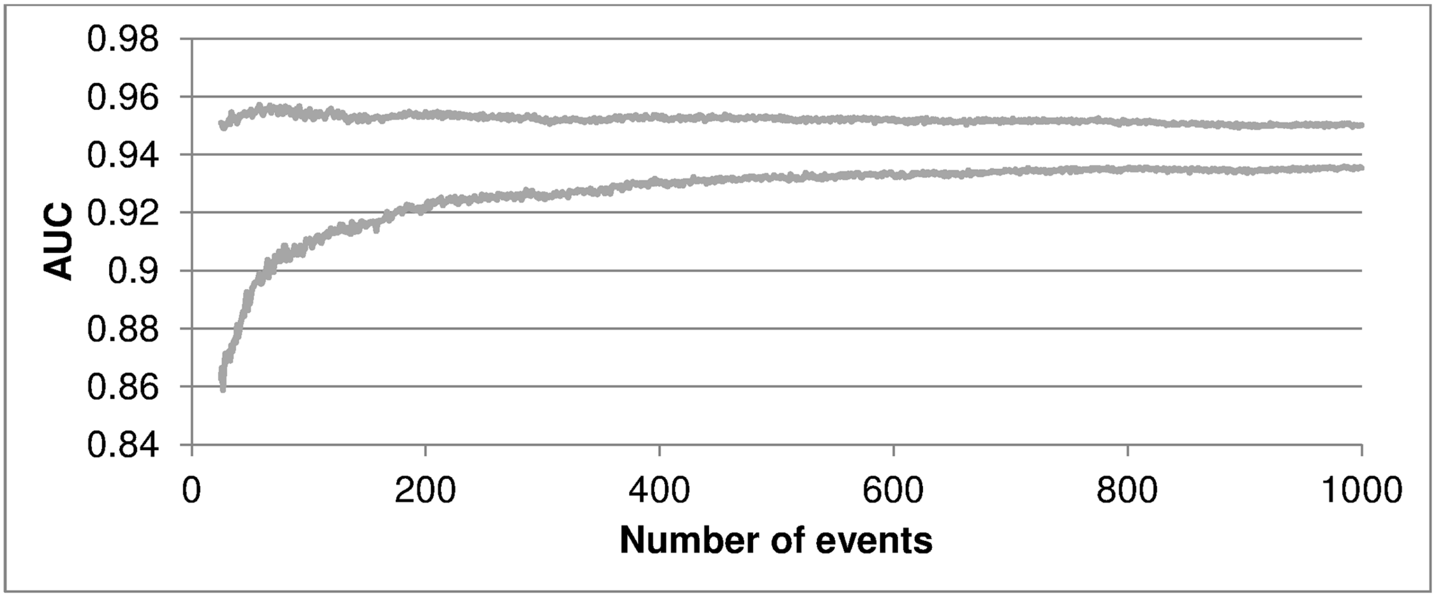

Fig 1. Confidence intervals for the area under the ROC curve according to the number of events in the sample. AUC, area under the ROC curve.

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.

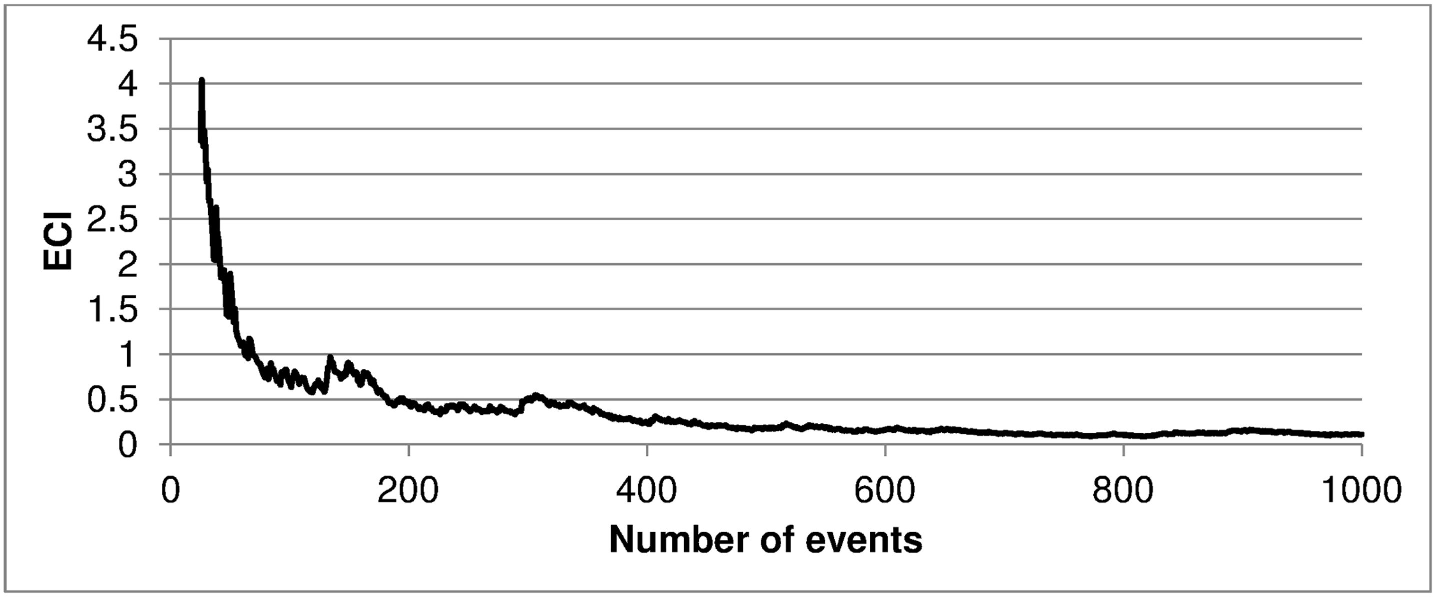

Fig 2. Estimated calibration index values according to the number of events in the sample. ECI, estimated calibration index.

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

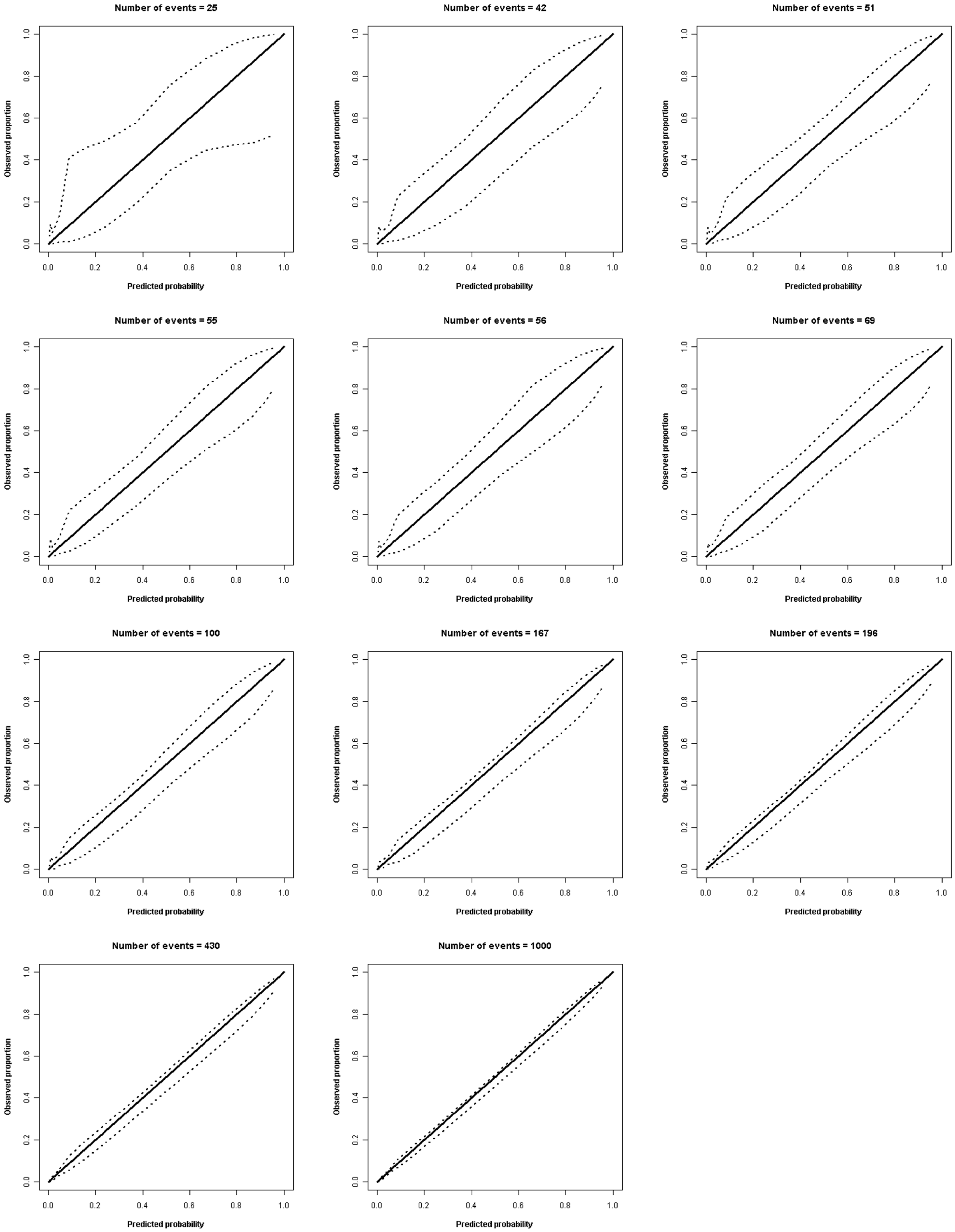

Fig 3. Smooth calibration plots (linear splines) for several sample sizes. The dashed lines denote the confidence intervals. The central line denotes the perfect prediction.

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