ELSEVIER

Journal of Clinical Epidemiology 143 (2022) 81–90

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Accuracy of approximations to recover incompletely reported logistic regression models depended on other available information

Toshihiko Takada[a,b], Jeroen Hoogland[a], Chris van Lieshout[a], Ewoud Schuit[a], Gary S. Collins[c], Karel G.M. Moons[a], Johannes B. Reitsma[a,*]

[a] *Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*
[b] *Department of General Medicine, Shirakawa Satellite for Teaching And Research (STAR), Fukushima Medical University, Fukushima, Japan*
[c] *Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK*

Accepted 24 November 2021; Available online 2 December 2021

Abstract

Objective: To provide approximations to recover the full regression equation across different scenarios of incompletely reported prediction models that were developed from binary logistic regression.

Study design and setting: In a case study, we considered four common scenarios and illustrated their corresponding approximations:

(A) Missing: the intercept, Available: the regression coefficients of predictors, overall frequency of the outcome and descriptive statistics of the predictors;

(B) Missing: regression coefficients and the intercept, Available: a simplified score;

(C) Missing: regression coefficients and the intercept, Available: a nomogram;

(D) Missing: regression coefficients and the intercept, Available: a web calculator.

Results: In the scenario A, a simplified approach based on the predicted probability corresponding to the average linear predictor was inaccurate. An approximation based on the overall outcome frequency and an approximation of the linear predictor distribution was more accurate, however, the appropriateness of the underlying assumptions cannot be verified in practice. In the scenario B, the recovered equation was inaccurate due to rounding and categorization of risk scores. In the scenarios C and D, the full regression equation could be recovered with minimal error.

Conclusion: The accuracy of the approximations in recovering the regression equation varied depending on the available information.

Keywords: Prediction model; Logistic regression; Equation; Intercept; Reverse engineering; Reporting

What is new?

- In reports where the intercept is not reported, the recovered intercept could be accurate; however, the appropriateness of underlying assumptions cannot be verified.
- In reports where only a simplified score is reported, the recovered prediction model equation is likely to be inaccurate as information is lost due to rounding and categorization that cannot be retrieved.
- With full knowledge of functional form of each predictor and interaction terms, it is possible to accurately recover the unreported regression equation in scenarios where tools are presented that can be used to estimate a probability (e.g., a nomogram or a web calculator).
- We propose approximations and guidance how to recover the logistic regression equation that is incompletely reported for various scenarios.

1. Introduction

Clinical prediction models can support decision-making by informing physicians, patients and their families on the probability of a health outcome [1–3]. Although prediction models can be derived by fitting machine learning tech-

Conflict of interests: GSC was supported by the NIHR Biomedical Research Centre, Oxford, and Cancer Research UK (programme grant: C49297/A27294). All other authors declared no competing interests.

* Corresponding author.
E-mail address: J.B.Reitsma-2@umcutrecht.nl (J.B. Reitsma).

https://doi.org/10.1016/j.jclinepi.2021.11.033

niques [4], when the outcome is binary, logistic regression modeling is most commonly used because of its ease of interpretation [5]. In a logistic regression based prediction model, an individual's risk of the outcome can be calculated using the individual's observed predictor values and the model's intercept and regression coefficients with the following equation:

$$LP = \log\left(\frac{p}{1-p}\right) = a + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k$$

$$p = \frac{\exp(LP)}{1 + \exp(LP)} \text{ or } p = \frac{1}{1 + \exp(-LP)}$$

where *LP* is the linear predictor; *p* is the predicted probability; a is the intercept; β_i is the regression coefficient of the predictor *i; X_i* is the observed value for the predictor *i; k* is the number of predictors in the model.

The Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement for prediction models clearly states that a prediction model's full regression equation should be reported (item 15a) [5,6]. However, incomplete reporting of prediction models is not uncommon, with some reviews showing incomplete reporting in about half of studies developing a prediction model [7–9]. The full equation of a prediction model is necessary when healthcare professionals use the model in practice or researchers evaluate the performance of the model as its original form in a different dataset (external validation) [10,11]. Sometimes, tools which can be used to estimate a probability for an individual (e.g., a nomogram or a web calculator) are reported, but the original regression equation is not available. In such situations, the model can be used for each individual, however, this process of manually entering data becomes too cumbersome when evaluating a large number of patients as in an external validation study.

Of course, the most obvious solution for an unreported full regression equation is to contact the authors of the paper and request the information. Our starting point is that this request was not successful. We aimed to provide approximations on how to recover the full regression equation in papers where the full regression equation is partially or completely missing. Such an attempt has been referred to as *reverse engineering* [12]. We considered scenarios which varied in (i) which component of the equation was missing and (ii) which other relevant pieces of information were available. For each scenario, we indicate a potential approximation to recover the missing information, and discuss its accuracy and limitations. We use a case study predicting the risk of having to undergo an operative delivery in laboring women.

2. Methods

2.1. Scenarios and approximations

We considered the following four common scenarios (Table 1).

Scenario (A)

It is common that the coefficient (or the odds ratio, where log(odds ratio) = coefficient) for each predictor is reported, but the intercept is not. In this scenario, regression coefficients, the overall frequency of the outcome, and descriptive statistics of each predictor can be used to recover the intercept. We consider two types of the approximation below.

Approximation of the intercept based on an estimated average linear predictor and the overall outcome frequency

A simple idea is to ignore the logistic link function and to assume that the mean of all model-based outcome probability predictions for individual patients corresponds to the risk of an "average individual" (i.e., a hypothetical individual who has the mean value for continuous variables and the proportions for categorical variables). This "average individual" can be achieved based on information that is usually available in a table of participant characteristics showing the frequency of categorical predictors as percentages and measures of central tendency (e.g., mean) and spread for continuous predictors (e.g., standard deviation) (Table 2).

Using this information, the intercept can be estimated as follows:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) \approx a \ + \ \beta 1 \ \bar{X}_1 + \ \beta 2 \ \bar{X}_2 \ + \ldots \beta_k \bar{X}_k$$

$$a \approx \log\left(\frac{\hat{p}}{1-\hat{p}}\right) - -\left(\beta_1 \bar{X}_1 \ + \beta \ _2 \bar{X}_2 \ + \ \ldots \beta_k \bar{X}_k\right)$$

where $\hat{p}$ is the overall frequency of the outcome; *a* is the intercept; β_i is the regression coefficient of the predictor *i; $\bar{X}_i$* is the mean/proportion value for the predictor *i*; and *k* is the number of predictors in the model.

While such an approximation would work in a linear model, it is generally flawed in other generalized linear models such as logistic regression. This is because they contain nonlinear link functions (e.g., the logit function for logistic regression). These nonlinear transformations cause the average of all individual risks to be generally unequal to the risk corresponding to the average linear predictor. Also, as a table of participant characteristics usually shows those descriptive statistics only for each original variable, this approximation is not applicable when the model includes non-linear or interaction terms. Further information on this approximation is available in the Supplementary material 1.

Approximation of the intercept based on the overall outcome frequency and an approximation of the linear predictor distribution

To improve upon the "average individual" approach, the idea is to approximate the covariate contributions and

Table 1. Missing and available information in each scenario

Scenario	Missing information	Available information
A	Intercept	Regression coefficients, overall outcome frequency and descriptive statistics of predictors
B	Intercept, regression coefficients	Simplified score
C	Intercept, regression coefficients	Nomogram
D	Intercept, regression coefficients	Web calculator

Table 2. Participant characteristics in the case study

	Spontaneous delivery	Operative delivery	Overall
	n = 4077 (71.9%)	n = 1590 (28.1%)	n = 5667
Maternal age, years, mean ± SD	32.0 ± 4.8	32.3 ± 4.7	32.0 ± 4.8
Gestational age, weeks, mean ± SD	40.1 ± 1.5	40.5 ± 1.4	40.2 ± 1.4
Birthweight, 100 g increments, mean ± SD	35.1 ± 5.1	36.2 ± 5.3	35.4 ± 5.2
Previous delivery, n (%)			
No	1990 (48.8)	1246 (78.4)	3236 (57.1)
Yes, but not by caesarean section	1596 (39.1)	119 (7.5)	1715 (30.3)
Yes, by caesarean section	491 (12.0)	225 (14.2)	716 (12.6)
Neonatal female gender, n (%)	1977 (48.5)	691 (43.5)	2668 (47.1)
Maternal diabetes, n (%)	120 (2.9)	49 (3.1)	169 (3.0)

SD = standard deviation

thereby provide information on the variability of the linear predictor. For simplicity, we assume a multivariate normal (MVN) distribution of the predictor distribution based on means and variances available in a table of participant characteristics. The remaining assumption is with respect to the correlation structure, for which we assume compound symmetry with an off-diagonal correlation ρ. The size of ρ needs to be chosen by the researcher. Based on this multivariate normal model, the estimated variance of the linear predictor $\hat{\sigma}^2$ then becomes

$$\hat{\sigma}^2 = \hat{\beta}\boldsymbol{S}\hat{\beta}^T$$

where $\hat{\beta}$ is a vector of the reported regression coefficients; and the $\boldsymbol{S}$ is the compound symmetric covariance matrix. The mean of the estimated linear predictor $\hat{\mu}_{lp}$ is

$$\hat{\mu}_{lp} = \hat{\mu}\hat{\beta}$$

where $\hat{\mu}$ is the vector containing the mean estimates for each predictor. Then using

$$\Pr(Y = 1) = \frac{1}{1 + e^{-a-\hat{\mu}\hat{\beta}-\sigma Z}}$$

where a is the intercept, and Z is a standard normally distributed random variable, the event rate or expected probability of $\Pr(Y = 1)$ equals

$$\mathbb{E}(PR(Y = 1|X)) = \int_{-\infty}^{+\infty}\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\frac{1}{1 + e^{-a-\hat{\mu}_{lp}-\sigma z}}\right)dz = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\left(\frac{e^{-\frac{z^2}{2}}}{1 + e^{-a-\hat{\mu}_{lp}-\sigma z}}\right)dz$$

This can be solved numerically for a if the overall outcome frequency is given.

The accuracy of the approximated intercept depends on the correct specification of the MVN approximation of the predictor distribution and the chosen value of ρ, which cannot be verified in practice. Nonetheless, the impact of the value of ρ could be examined by varying its value over a sensible range. The range of approximated intercepts across these sensitivity analyses then has to lie within reasonable bounds for the problem at hand. The R code for this approximation is available in the supplementary material 2.

Scenario (B)

In this scenario, neither the regression coefficients nor the intercept is reported, but a simplified score, for example Table 3, is available.

While there are various ways to derive a simplified score, it is often done by converting the coefficients for each predictor to integers (e.g., dividing all regression coefficients by the smallest regression coefficient and rounding to the whole number) [5]. This is common, but there have been better approaches for deriving a simplified score [13,14]. Continuous variables are often categorized into two or more groups resulting in a loss of precision. The total sum of all the predictor integers for an individual, referred to as the score, can then categorized into risk groups together with their group's corresponding observed outcome frequencies and/or the model's mean predicted probabilities [15]. Although a simplified score is easy to use, it loses predictive information due to rounding and

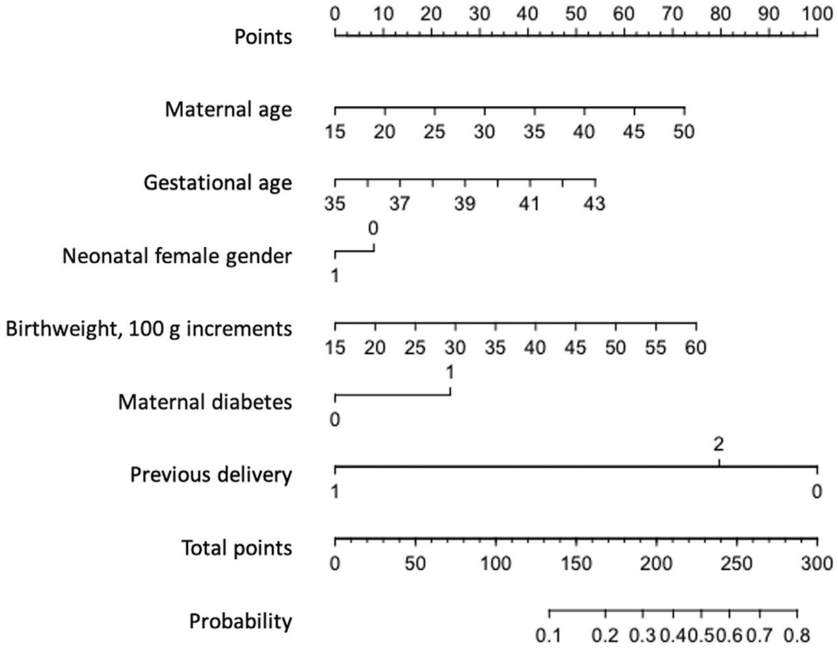


Fig. 1. Nomogram developed for the case study.

Table 3. Simplified score developed for the case study

Variable	Score	
Maternal age	1	
Gestational age	4	
Previous delivery (Reference: None)		
Not by caesarean section	-60	
By caesarean section	-12	
Neonatal female gender	-5	
Birthweight, 100 g increments	1	
Maternal diabetes	14	
Score	Mid-point*	Mean predicted probability
129-204	166.5	8%
204-224	214	30%
224-264	244	46%

* A mid-point is the average of the lower and upper ranges in each score group. Although it is not usually shown in the scoring system, it is calculated here to use in the approximation for recovering the regression equation.

categorization. One can try to estimate the regression coefficients and the intercept using the limited information available from a reported simplified score as follows:

$$T \approx \frac{(\beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k)}{w}$$

$$\beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k \approx wT$$

$$\log\left(\frac{p'}{1-p'}\right) \approx a + wT$$

where T is the total score; β_i is the regression coefficient of the predictor i; X_i is the observed value for the predictor i; k is the number of predictors in the model; w is the value used to divide each regression coefficient (i.e., min (β_1, β_2, β_3,..β_k); p' is the mean predicted probability for each value of T or categorizations of T; and a is the intercept.

When the total score T has been categorized, the corresponding score for an individual who has the mean predicted probability in each score group is rarely available. The alternative is to use the mean score in each score group. However, this is also not commonly reported. Instead, one can calculate the mid-point in each score group (i.e., the average of the lower and upper ranges in each score group), but it is worth noting that the difference between the mid-point and the score corresponding to the mean predicted probability in each score group is unlikely to be aligned. The available data then consists of rows for each score group stating the mid-point and the corresponding mean predicted probability (p'). The relation between these quantities can be written as a logistic model (see equation above), which can then be used to estimate w and a. Each coefficient is approximated using w and the score assigned to each predictor (s_j) as below.

$$\beta_j \approx w s_j$$

Scenario (C)

In this scenario, no information on the regression equation is reported, but a nomogram (Fig. 1) has been available.

A nomogram is a graphical presentation of a prediction model, which enables estimation of a predicted probability for an individual without a calculator [16]. In Fig. 1, the top bar shows how many points a certain value of each predictor (in the bars below) corresponds to. Summing the individual points for each predictor, results in a total points score. The bottom two bars show total points and corresponding predicted probabilities.

Under certain conditions, it is possible to reconstruct the original equation if a nomogram is provided. These conditions include knowledge of (i) functional form of each predictor (e.g., categorical, continuous linear or restricted cubic spline function for non-linear terms), and (ii) interaction terms (i.e., which and how predictors were included in interaction terms). The details of the process are explained in the supplementary material 3. For example, in a simple model which does not include any non-linear or interaction terms, first, measure the length of the bar for "total points". For precise measurement, we recommend to use a software which enables digitizing the nomogram. Then, calculate the distance for 1 point. Second, transform the probability on the bar "Probability" into linear predictors, by logit transformation. Again, measure the length of the bar "Probability", and calculate the distance for 1 in the linear predictor. Then, the total points score which corresponds to 1 in the linear predictor can be calculated. Third, measure the length of the bar for each predictor. Then, the coefficient can be calculated as how much linear predictor increases when one-unit changes in each predictor. Finally, the intercept can be estimated by calculating a predicted probability for a certain individual using the coefficients calculated through the above process using the equation below.

$$a = \log\left(\frac{p}{1-p}\right) - \left(\beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k\right)$$

where a is the intercept; p is a predicted probability for a certain individual; β_i is the predictor weight or regression coefficient of the predictor i; X_i is the value of a certain individual for the predictor i; and k is the number of predictors in the model.

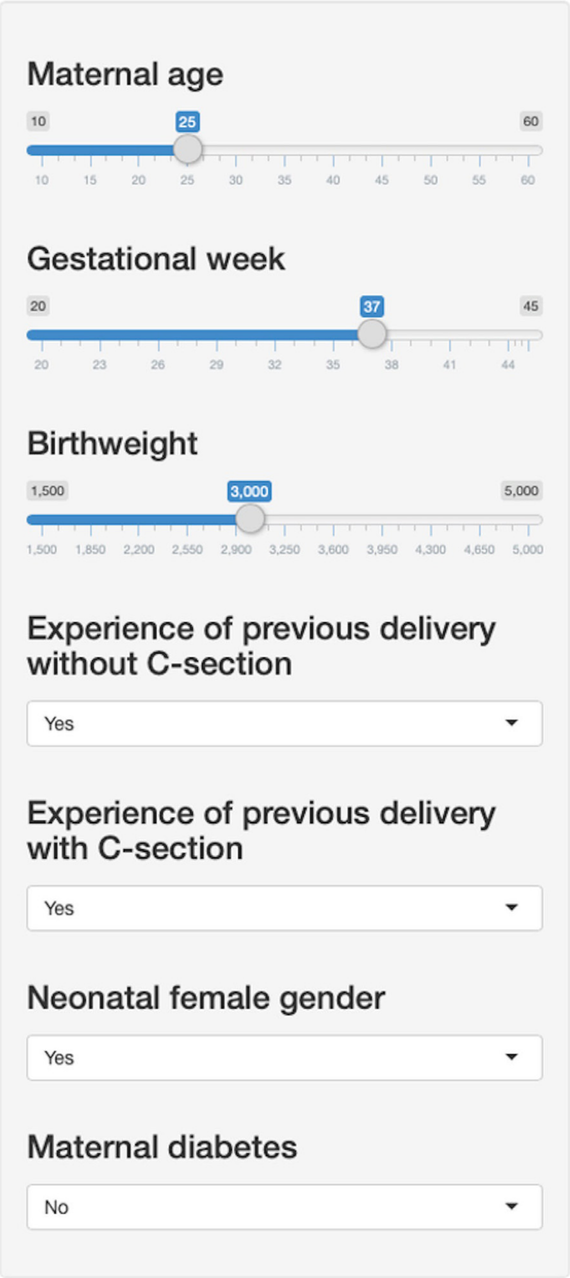
Scenario (D)

In this scenario, the regression equation has not been reported, but a link to a web calculator is available (Fig. 2).

To use the model in new individuals, users can enter the values of the predictors on the website, and the web calculator produces an estimate of the predicted probability for that individual. Similar to the scenario (C), it is possible to reconstruct the original equation from a web calculator if full information about functional form of each predictor and interaction terms is available. When non-linear or interaction terms are not used in a model, reverse engineering starts by first entering values of the predictors for a particular individual (individual 1) into the web calculator to obtain his/her predicted outcome probability (p_1). In this step, the values of the predictors can be arbitrarily chosen. For the next individual, change the value of one predictor (X_i), while leaving the remaining predictor values fixed (individual 2), and obtain the predicted probability for individual 2 (p_2). The coefficient β_i for X_i can be calculated as the difference of linear predictors in patient 1 and 2 divided by the change in X_i (X_{diff}) as following:

$$\frac{LP1 - LP2}{X_{diff}} = \frac{\log\left(\frac{p1}{1-p1}\right) - \log\left(\frac{p2}{1-p2}\right)}{X_{diff}} = \beta i$$

where *LP1* and *LP2* are the linear predictor in individual 1 and 2, respectively; X_{diff} is the change in X_i; *p1* and *p2* are the predicted probability in individual 1 and 2, respectively; β_i is the predictor weight or regression coefficient of the predictor *i*.



Maternal age

25

Gestational week

37

Birthweight

3,000

Experience of previous delivery without C-section

Yes

Experience of previous delivery with C-section

Yes

Neonatal female gender

Yes

Maternal diabetes

No

Fig. 2. Image of the web calculator developed for the case study The web calculator is available from https://pred-model.shinyapps.io/reverse_engineering/.

Preferably, the difference in X_i between individuals one and two is taken to be large to avoid imprecision due to rounding errors of the provided predicted probabilities. By repeating this process for each predictor, the predictor coefficients (β_1, β_2, β_3... β_k) can be calculated. Finally, the intercept (*a*) can be obtained by using these recovered predictor coefficients and arbitrary values for each predictor as follows:

$$LP1 \; = \log\left(\frac{p1}{1-p1}\right) = a + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k$$

$$a \; = \log\left(\frac{p1}{1-p1}\right) - (\beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k)$$

2.2. Case study

To illustrate the scenarios and the corresponding approximations, we used an existing dataset comprising 5667 laboring women with a singleton term pregnancy in cephalic presentation, previously used to develop a multinomial model to predict the mode of delivery [17]. For illustrative purposes, the categorical outcome of the original publication was dichotomized as operative delivery (i.e., instrumental vaginal delivery or caesarean section) vs. spontaneous vaginal delivery. Here, we focused on six antepartum predictors being maternal age (years), gestational age (weeks), birthweight (100g increments), previous delivery (0 = none, 1 = yes, but not by caesarean section, 2 = yes, by caesarean section), neonatal gender (0 = male, 1 = female), and maternal diabetes (0 = no, 1 = yes). To illustrate all of the approximations above, all continuous variables were modeled linearly and no interaction term was included. In the current analysis, we used one of the imputed datasets used in the original analysis.

2.3. Comparison between the original and the recovered regression equation in each scenario

The original logistic regression model equation was taken to represent the reference model. The recovered regression equation in each scenario was compared to that of the reference model. To assess the impact of the recovered equation on predicted probabilities, a predicted probability was estimated using the recovered equation in individuals with 10th, 50th, and 90th percentile predicted probabilities estimated by the reference model (i.e., low, medium, and high risk, respectively).

All analyses were performed using R 3.6.1. [18].

3. Results

3.1. Case study

Participant characteristics are shown in Table 2. Among 5667 laboring women, 1590 (28.1%) underwent an operative delivery. The reference values of the intercept, coefficients for each predictor, and the model performance measures are shown in Table 4.

Scenario (A)

Approximation of the intercept based on an estimated average linear predictor and the overall outcome frequency

To recover the intercept, we assumed that the predicted probability for the patient who had the mean/proportion value of each predictor ("average individual") would be close to the observed proportion of the outcome of 28.1%. By using the mean/proportion value of each predictor in Table 2 and the value of coefficients in Table 4, the following equation was obtained.

$$\begin{aligned}\log\left(\frac{0.281}{1-0.281}\right) &\approx a \; + \; \beta mat_{age\bar{X}_{matage}} + \beta ges\_age \\ &+ \bar{X}_{ges\_age} + \beta weight\bar{X}_{weight} \\ &+ \beta no\_CS\bar{X}_{no\_CS} + \beta CS\bar{X}_{CS} + \beta gen\bar{X}_{gen} \\ &+ \ldots \beta diabe\bar{X}_{diabe} \\ &\approx a + 0.048 \;\times\; 32.0 + 0.157 \;\times\; 40.2 + (-2.323) \\ &\times\; 0.303 + (-0.473) \times\; 0.126 \\ &+ (-0.186) \times\; 0.471 + 0.039 \;\times\; 35.4 + 0.555 \;\times\; 0.03\end{aligned}$$

where *a* is the intercept; each β_i represents the predictor weight or regression coefficient; each $\bar{X}_i$ is the mean/proportion value of each predictor; *mat_age* stands for maternal age, *ges_age* for gestational age, *weight* for birthweight, *no CS* for previous delivery not by caesarean section, *CS* for previous delivery by caesarean section, *gen* for neonatal female gender, and *diabe* for maternal diabetes.

Then, the intercept was calculated as -9.333, while the reference value was -9.563. The estimated predicted probabilities based on the reference model and the recovered equation were 0.054 vs. 0.068, 0.309 vs. 0.366, and 0.484 vs. 0.549 for low, medium and high-risk patients, respectively.

Approximation of the intercept based on the overall outcome frequency and an approximation of the linear predictor distribution

When we set the off-diagonal correlation ρ as 0.3 (i.e., the assumed correlation between all predictors), the estimated intercept was -9.525, while the reference value was -9.563. The estimated predicted probabilities based on the reference model and the recovered equation were 0.054 vs. 0.057, 0.309 vs. 0.323, and 0.484 vs. 0.501, respectively. When the value of ρ was changed between 0 and 1, the recovered intercept ranged from -9.573 to -9.400.

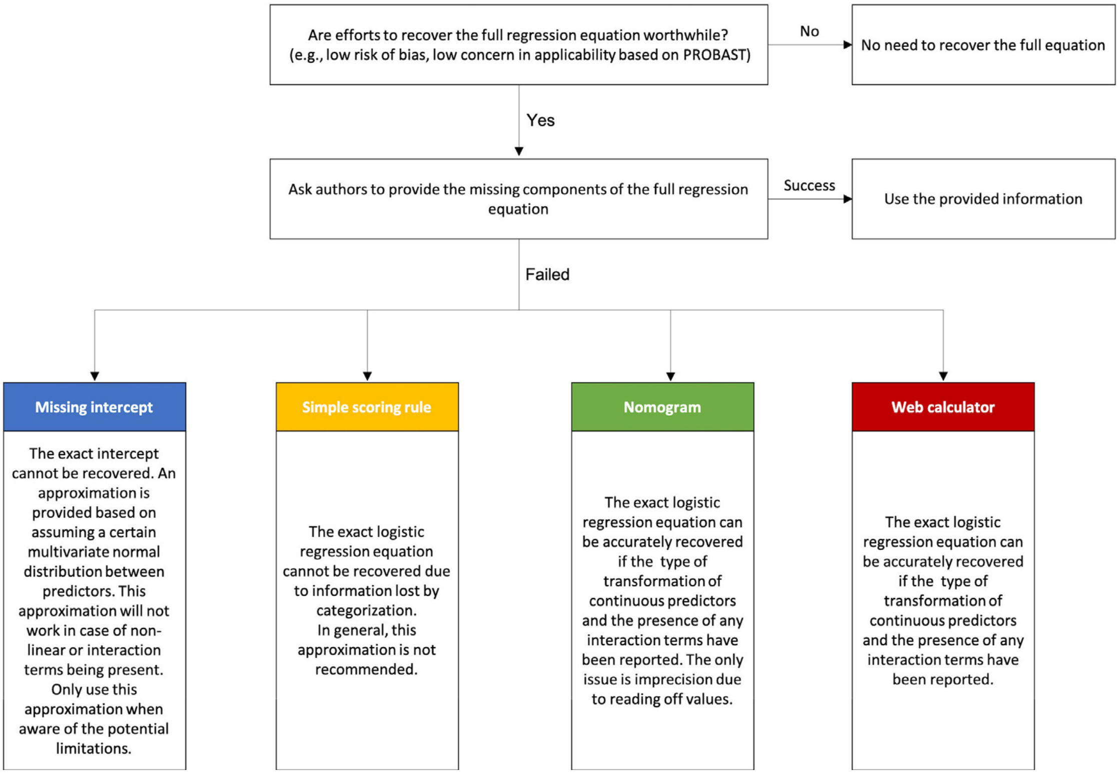
Scenario (B)

A simplified score derived from the prediction model is shown in Table 3. The score for each predictor was derived by dividing each coefficient by the smallest one for birthweight (0.039) and rounded to an integer. In this example, we made the score with three groups categorized based on tertiles of the total score and their corresponding mean predicted probabilities. Based on the mid-point and

Table 4. Reference and recovered values in each scenario of the regression formula and estimated probabilities

		Scenario (A) Missing intercept		**Scenario (B) Simplified score**	**Scenario (C) Nomogram**	**Scenario (D) Web calculator**
	Reference values	**Approximation 1***	**Approximation 2†**			
Regression formula						
Intercept	-9.563	-9.333	-9.525	-7.098	-9.580	-9.552
Maternal age	0.048	NA	NA	0.029	0.048	0.048
Gestational age	0.157	NA	NA	0.116	0.157	0.156
Previous delivery (Reference: None)						
Not by caesarean section	-2.323	NA	NA	-1.740	-2.326	-2.318
By caesarean section	-0.473	NA	NA	-0.348	-0.473	-0.468
Neonatal female gender	-0.186	NA	NA	-0.145	-0.185	-0.184
Birthweight, 100 g increments	0.039	NA	NA	0.029	0.039	0.039
Maternal diabetes	0.555	NA	NA	0.406	0.559	0.553
Estimated probability						
Probability for a low risk patient	0.054	0.068	0.057	0.129	0.054	0.054
Probability for a medium risk patient	0.309	0.366	0.323	0.310	0.311	0.308
Probability for a high risk patient	0.484	0.549	0.501	0.380	0.487	0.482

* Approximation 1 is based on an estimated average linear predictor and the overall outcome frequency.
† Approximation 2 is based on the overall outcome frequency and an approximation of the linear predictor distribution.



Are efforts to recover the full regression equation worthwhile? (e.g., low risk of bias, low concern in applicability based on PROBAST)

No need to recover the full equation

Ask authors to provide the missing components of the full regression equation

Use the provided information

The exact intercept cannot be recovered. An approximation is provided based on assuming a certain multivariate normal distribution between predictors. This approximation will not work in case of non-linear or interaction terms being present. Only use this approximation when aware of the potential limitations.

The exact logistic regression equation cannot be recovered due to information lost by categorization. In general, this approximation is not recommended.

The exact logistic regression equation can be accurately recovered if the type of transformation of continuous predictors and the presence of any interaction terms have been reported. The only issue is imprecision due to reading off values.

The exact logistic regression equation can be accurately recovered if the type of transformation of continuous predictors and the presence of any interaction terms have been reported.

Fig. 3. Recommendation when having a published prediction model in which the regression equation is incompletely reported.

mean predicted probability, the following equations could be derived for each group:

$$\log\left(\frac{0.08}{1-0.08}\right) \approx a + 166.5 \quad \times w$$

$$\log\left(\frac{0.3}{1-0.3}\right) \approx a + 214 \quad \times w$$

$$\log\left(\frac{0.46}{1-0.46}\right) \approx a + 244 \quad \times w$$

where a = intercept, w = the value which was used to divide each coefficient.

These equations can be simultaneously solved for *a* and *w* using logistic regression modelling, resulting in the values of -7.098 and 0.029, respectively. Then, each regression coefficient was estimated using *w* and the score assigned to each predictor. For instance, the coefficient for gestational age was calculated as follows:

$$4 \times 0.029 = 0.116$$

As shown in Table 4, the recovered equation and the estimated predicted probabilities were very different from the reference values.

Scenario (C)

The nomogram derived from this case study is shown in Fig. 1. The detail of the procedure to recover the regression equation in this scenario is explained in supplementary material 3. As shown in Table 4, the recovered coefficients and intercept were precisely estimated. Accordingly, the predicted probability in low, medium, and high-risk patients were also correctly estimated.

Scenario (D)

A web calculator for the developed prediction model is available at https://pred-model.shinyapps.io/reverse_engineering/. The coefficient for each predictor was obtained as the difference in linear predictors divided by the change in a certain predictor (while keeping the values of the other predictors constant). For example, the web calculator showed the predicted probability of 1.2% for an individual with maternal age of 25, gestational week of 37, birthweight of 3000, experience of previous delivery with and without C-section, neonatal female gender, and no maternal diabetes. When changing the maternal age from 25 to 60, the predicted probability increased to 6.2%. Then, the following equations were derived to determine the regression coefficient of maternal age ($\beta_{\mathrm{mat\_age}}$).

$$\log\left(\frac{0.012}{1-0.012}\right) = a + 25 \times \beta_{mat\_age} + 37 \times \beta_{ges\_age} + 30 \times \beta_{\ weight} + 1 \times \beta_{\ no\_CS} + 1 \times \beta_{\ CS} + 1 \times \beta_{gen} + 0 \times \beta_{diabe}$$

$$\log\left(\frac{0.062}{1-0.062}\right) = a + 60 \times \beta_{\ mat\_age} + 37 \times \beta_{ges\_age} + 30 \times \beta_{weight} + 1 \times \beta_{no\_CS} + 1 \times \beta_{CS} + 1 \times \beta_{gen} + 0 \times \beta_{diabe}$$

Then,

$$\log\left(\frac{0.062}{1-0.062}\right) - \log\left(\frac{0.012}{1-0.012}\right) = 35 \times \beta_{mat\_age}$$

and β_{mat_age} was calculated as 0.048, which was identical to that of the reference model. The coefficients for the other predictors were estimated in the same way. Finally, with the estimated coefficients, the intercept could be obtained by calculating the predicted probability in a certain individual. Similar to the scenario (C), the regression equation and the predicted probability were accurately estimated.

4. Discussion

We discussed and illustrated the accuracy of various approximations for recovering the full regression equation for incompletely reported prediction models developed with logistic regression.

In the most common situation where the coefficient/odds ratios for each predictor is reported, but the intercept is not (scenario (A)) [19], external validation of the prediction model in its original form is impossible. If a separate dataset of new individuals from the target population is available, the intercept can be re-estimated (recalibration in the large) by fitting a logistic regression model using the linear predictor calculated from the available coefficients/odd ratios without the intercept as an offset (i.e., coefficient fixed as 1). However, the estimated intercept is then tailored to the validation dataset and is likely to be different from the (unreported) intercept of the original prediction model. Obviously, this approach is possible only when there is available dataset that is sufficiently large [20]. When the model is intended to be used for prediction of the outcome in an individual, it is impossible to use the model without the intercept. In this scenario, we suggested two types of the approximation which can be used when the model does not include non-linear or interaction terms. In the simplified approach based on the "average individual", the intercept was not recovered accurately due to substantial bias caused by clear invalidity of assumptions. On the other hand, the approximation based on the overall outcome frequency and an approximation of the linear predictor distribution showed good accuracy of the recovered intercept. This approximation assumed a MVN distribution of the predictors with a compound symmetric correlation structure. Sensitivity analyses with respect to the assumed correlation structure can easily be performed and provide information on the robustness of the approximated intercept. Nonetheless, sensitivity to deviations from multivariate normality is hard to assess. Thus, this approximation should be used only when the readers are aware of its limitations and the potential risk of bias.

When a simplified score is presented without the original regression equation (scenario (B)), we found that it is impossible to recover the precise regression equation due to the lost information by rounding and categorization. The risk of inaccurate approximation increases with fewer/broader risk categories. Also, dissociation between the mid-point and the corresponding mean predicted probability in that score group is often problematic. In general, we do not recommend this approximation. Although a simplified score is commonly presented for the sake of ease of use, researchers should always present the original regression equation as well to allow others to properly validate and update (if needed) the underlying prediction model [13].

In the scenarios where a web calculator or a nomogram are presented, the full regression equation can be accurately recovered even when non-linear and/or interaction terms are included as long as full knowledge of functional form of each predictor and interaction terms is available. Yet, it is not very likely that such detailed information is available when the full regression equation is not properly reported. The approximations can be imprecise when reading values from the nomogram is difficult.

Possible explanations for the incomplete reporting of model's regression equation may be that researchers are just unaware of its importance, or that one intentionally hides the equation to protect intellectual property or charge royalty for the use of a prediction model. After the introduction of the TRIPOD statement in 2015, information necessary for individual risk prediction was more frequently reported in studies published in high-impact journals, but still not sufficient (42% between 2016 and 2017 vs. 27% between 2012 and 2014) [21]. In addition, Prediction model Risk of Bias Assessment Tool (PROBAST) has been published in 2019 [22,23]. In this risk of bias assessment tool for prediction model studies, it is clearly stated that the full regression equation of the developed model should be fully reported to allow others to correctly apply the model to other individuals. It is expected that full regression equations will be reported more frequently by disseminating these guidelines. However, the experience with other reporting guidelines is that improving adherence is a slow process requiring continuous attention and efforts [24,25].

To summarize, we propose the following guidance when faced with validating or implementing a published logistic prediction model in which the regression equation is incompletely reported (Fig. 3). First, evaluate whether the model is valid and valuable enough to make efforts to recover the full regression equation. The risk of bias and its applicability to one's clinical question can be assessed by using PROBAST. Second, before trying to recover the regression equation, one should first ask the authors to provide their missing information. If this fails, one can start the process of recovering the regression equation based on the information and warnings provided in this paper.

Contributors

All authors provided a substantial contribution to the conceptualization, design, interpretation of the case study, and writing the draft. TT performed the analyses of the case study and wrote the first draft. CVL, ES, KGMM, and JBR initiated this project. JH and GSC led the methodology that underpins the methods in this article.

Funding

JH and JBR were supported by a TOP grant of the Netherlands Organisation for Health Research and Development (grant: 91215058).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jclinepi.2021.11.033.

References

[1] Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 2012;98:683–90.

[2] Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:b375.

[3] 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.

[4] Zhang Z, Liu J, Xi J, Gong Y, Zeng L, Ma P. Derivation and validation of an ensemble model for the prediction of agitation in mechanically ventilated patients maintained under light sedation. Crit Care Med 2021;49:e279–ee90.

[5] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1–73.

[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.

[7] Baart SJ, Dam V, Scheres LJJ, Damen J, Spijker R, Schuit E, et al. Cardiovascular risk prediction models for women in the general population: A systematic review. PLoS One 2019;14:e0210329.

[8] Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 2016;353:i2416.

[9] Kleinrouweler CE, Cheong-See FM, Collins GS, Kwee A, Thangaratinam S, Khan KS, et al. Prognostic models in obstetrics: available, but far from applicable. Am J Obstet Gynecol 2016;214:79–90 e36.

[10] Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.

[11] Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009;338:b606.

[12] Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J Am Med Inform Assoc 2019;26:1651–4.

[13] Bonnett LJ, Snell KIE, Collins GS, Riley RD. Guide to presenting clinical prediction models for use in clinical settings. BMJ 2019;365:l737.

[14] 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.

[15] Rietveld RP, ter Riet G, Bindels PJ, Sloos JH, van Weert HC. Predicting bacterial cause in infectious conjunctivitis: cohort study on informativeness of combinations of signs and symptoms. BMJ 2004;329:206–10.

[16] Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. J Clin Oncol 2008;26:1364–70.

[17] Schuit E, Kwee A, Westerhuis ME, Van Dessel HJ, Graziosi GC, Van Lith JM, et al. A clinical prediction model to assess the risk of operative delivery. BJOG 2012;119:915–23.

[18] R Core Team R: A language and environment for statistical computing, Vienna, Austria: R Foundation for Statistical Computing; 2021. URL https://www.R-project.org/.

[19] Heus P, Damen J, Pajouheshnia R, Scholten R, Reitsma JB, Collins GS, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. BMC Med 2018;16:120.

[20] Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE Jr, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 2019;38:1276–96.

[21] Zamanipoor Najafabadi AH, Ramspek CL, Dekker FW, Heus P, Hooft L, Moons KGM, et al. TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models. BMJ Open 2020;10:e041537.

[22] Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170:W1–W33.

[23] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170:51–8.

[24] Dechartres A, Trinquart L, Atal I, Moher D, Dickersin K, Boutron I, et al. Evolution of poor reporting and inadequate methods over time in 20 920 randomised controlled trials included in Cochrane reviews: research on research study. BMJ 2017;357:j2490.

[25] Korevaar DA, Wang J, van Enst WA, Leeflang MM, Hooft L, Smidt N, et al. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD. Radiology 2015;274:781–9.