Calculating the sample size required for developing a clinical prediction model

Clinical prediction models aim to predict outcomes in individuals, to inform diagnosis or prognosis in healthcare. Hundreds of prediction models are published in the medical literature each year, yet many are developed using a dataset that is too small for the total number of participants or outcome events. This leads to inaccurate predictions and consequently incorrect healthcare decisions for some individuals. In this article, the authors provide guidance on how to calculate the sample size required to develop a clinical prediction model.

Richard D Riley professor of biostatistics¹, Joie Ensor lecturer in biostatistics¹, Kym E Snell lecturer in biostatistics¹, Frank E Harrell Jr professor of biostatistics², Glen P Martin lecturer in health data sciences³, Johannes B Reitsma associate professor⁴, Karel G M Moons professor of clinical epidemiology⁴, Gary Collins professor of medical statistics⁵, Maarten van Smeden assistant professor⁴ ⁵ ⁶

¹Centre for Prognosis Research, School of Primary, Community and Social Care, Keele University, Staffordshire ST5 5BG, UK; ²Department of Biostatistics, Vanderbilt University School of Medicine, Nashville TN, USA; ³Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK; ⁴Julius Center for Health Sciences, University Medical Center Utrecht, Utrecht, Netherlands; ⁵Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK; ⁶Department of Clinical Epidemiology, Leiden University Medical Center Leiden, Netherlands

Summary points

Patients and healthcare professionals require clinical prediction models to accurately guide healthcare decisions

Larger sample sizes lead to the development of more robust models

Data should be of sufficient quality and representative of the target population and settings of application

It is better to use all available data for model development (ie, avoid data splitting), with resampling methods (such as bootstrapping) used for internal validation

When developing prediction models for binary or time-to-event outcomes, a well known rule of thumb for the required sample size is to ensure at least 10 events for each predictor parameter

The actual required sample size is, however, context specific and depends not only on the number of events relative to the number of candidate predictor parameters but also on the total number of participants, the outcome proportion (incidence) in the study population, and the expected predictive performance of the model

We propose to use such information to tailor sample size requirements to the specific setting of interest, with the aim of minimising the potential for model overfitting while targeting precise estimates of key parameters

Our proposal can be implemented in a four step procedure and is applicable for continuous, binary, or time-to-event outcomes

The pmsampsize package in Stata or R allows researchers to implement the procedure

Clinical prediction models are needed to inform diagnosis and prognosis in healthcare.¹⁻⁴ Well known examples include the Wells score,⁵ QRISK,⁶ and the Nottingham prognostic index.⁷ Such models allow health professionals to predict an individual’s outcome value, or to predict an individual’s risk of an outcome being present (diagnostic prediction model) or developed in the future (prognostic prediction model). Most prediction models are developed using a regression model, such as linear regression for continuous outcomes (eg, pain score), logistic regression for binary outcomes (eg, presence or absence of pre-eclampsia), or proportional hazards regression models for time-to-event data (eg, recurrence of venous thromboembolism).⁸ An equation is then produced that can be used to predict an individual’s outcome value or outcome risk conditional on his or her values of multiple predictors, which might include basic characteristics such as age, weight, family history, and comorbidities; biological measurements such as blood pressure and biomarkers; and imaging or other test results. Supplementary material S1 shows examples of regression equations.

Developing a prediction model requires a development dataset, which contains data from a sample of individuals from the target population, containing their observed predictor values (available at the intended moment of prediction¹¹) and observed outcome.
The sample size of the development dataset must be large enough to develop a prediction model equation that is reliable when applied to new individuals in the target population. What constitutes an adequately large sample size for model development is, however, unclear, with various blanket “rules of thumb” proposed and debated. This has created confusion about how to perform sample size calculations for studies aiming to develop a prediction model.

In this article we provide practical guidance for calculating the sample size required for the development of clinical prediction models, which builds on our recent methodology papers. We suggest that current minimum sample size rules of thumb are too simplistic and outline a more scientific approach that tailors sample size requirements to the specific setting of interest. We illustrate our proposal for continuous, binary, and time-to-event outcomes and conclude with some extensions.

Moving beyond the 10 events per variable rule of thumb

In a development dataset, the effective sample size for a continuous outcome is determined by the total number of study participants. For binary outcomes, the effective sample size is often considered equal to the minimum of the number of events (those with the outcome) and non-events (those without the outcome); time-to-event outcomes are often considered roughly equal to the total number of events. When developing prediction models for binary or time-to-event outcomes, an established rule of thumb for the required sample size is to ensure at least 10 events for each predictor parameter (ie, each β term in the regression equation) being considered for inclusion in the prediction model equation. This is widely referred to as needing at least 10 events per variable (10 EPP). The word “variable” is, however, misleading as some predictors actually require multiple β terms in the model equation—for example, two β terms are needed for a categorical predictor with three categories (eg, tumour grades I, II, and III), and two or more β terms are needed to model any non-linear effects of a continuous predictor, such as age or blood pressure. The inclusion of interactions between two or more predictors also increases the number of model parameters. Hence, as prediction models usually have more parameters than actual predictors, it is preferable to refer to events per candidate predictor parameter (EPP). The word candidate is important, as the amount of model overfitting is dictated by the total number of predictor parameters considered, not just those included in the final model equation.

The rule of at least 10 EPP has been widely advocated perhaps as a result of its simplicity, and it is regularly used to justify sample sizes within published articles, grant applications, and protocols for new model development studies, including by ourselves previously. The most prominent work advocating the rule came from simulation studies conducted in the 1990s, although this work actually focused more on the bias and precision of predictor effect estimates than on the accuracy of risk predictions from a developed model. The adequacy of the 10 EPP rule has often been debated. Although the rule provides a useful starting point, counter suggestions include either lowering the EPP to below 10 or increasing it to 15, 20, or even 50. These inconsistent recommendations reflect that the required EPP is actually context specific and depends not only on the number of events relative to the number of candidate predictor parameters but also on the total number of participants, the outcome proportion (incidence) in the study population, and the expected predictive performance of the model. This finding is unsurprising as sample size considerations for other study designs, such as randomised trials of interventions, are all context dependent and tailored to the setting and research question. Rules of thumb have also been advocated in the continuous outcome setting, such as two participants per predictor, but these share the same concerns as for 10 EPP.

Sample size calculation to ensure precise predictions and minimise overfitting

Recent work by van Smeden et al and Riley et al describe how to calculate the required sample size for prediction model development, conditional on the user specifying the overall outcome risk or mean outcome value in the target population, the number of candidate predictor parameters, and the anticipated model performance in terms of overall model fit (R²). These authors’ approaches can be implemented in a four step procedure. Each step leads to a sample size calculation, and ultimately the largest sample size identified is the one required. We describe these four steps, and, to aid general readers, provide the more technical details of each step in the figures.

Step 1: What sample size will produce a precise estimate of the overall outcome risk or mean outcome value?

Fundamentally, the sample size must allow the prediction model’s intercept to be precisely estimated, to ensure that the developed model can accurately predict the mean outcome value (for continuous outcomes) or overall outcome proportion (for binary or time-to-event outcomes). A simple way to do this is to calculate the sample size needed to precisely estimate (within a small margin of error) the intercept in a model when no predictors are included (the null model). Figure 1 shows the calculation for binary and time-to-event outcomes, and we generally recommend aiming for a margin of error of ≤0.05 in the overall outcome proportion estimate. For example, with a binary outcome that occurs in half of individuals, a sample size of at least 385 people is needed to target a confidence interval of 0.45 to 0.55 for the overall outcome proportion, and thus an error of at most 0.05 around the true value of 0.5. To achieve the same margin of error with outcome proportions of 0.1 and 0.2, at least 139 and 246 participants, respectively, are required.

For time-to-event outcomes, a key time point needs to be identified, along with the anticipated outcome event rate. For example, with an anticipated event rate of 10 per 100 person years of the entire follow-up, the sample size must include a total of 2366 person years of follow-up to ensure an expected margin of error of ≤0.05 in the estimate of a 10 year outcome probability of 0.63, such that the expected confidence interval is 0.58 to 0.68.

For continuous outcomes, the anticipated mean and variance of outcome values must be prespecified, alongside the anticipated percentage of variation explained by the prediction model (see supplementary material S2 for details).

Step 2: What sample size will produce predicted values that have a small mean error across all individuals?

In addition to predicting the average outcome value precisely (see step 1), the sample size for model development should also aim for precise predictions across the spectrum of predicted values. For binary outcomes, van Smeden et al use simulation across a wide range of scenarios to evaluate how the error of predicted outcome probabilities from a developed model...
depends on various characteristics of the development dataset sampled from a target population. They found that the number of candidate predictor parameters, total sample size, and outcome proportion were the three main drivers of a model’s mean predictive accuracy. This led to a sample size formula (fig 2) to help ensure that new prediction models will, on average, have a small prediction error in the estimated outcome probabilities in the target population (as measured by the mean absolute prediction error, MAPE). The calculation requires the number of candidate predictor parameters and the anticipated outcome proportion in the target population to be prespecified. For example, with 10 candidate predictor parameters and an outcome proportion of 0.3, a sample size of at least 461 participants and 13.8 EPP is required to target a mean absolute error of 0.05 between observed and true outcome probabilities (see fig 2 for calculation). The calculation is available as an interactive tool (https://mvansmeden.shinyapps.io/BeyondEPV/) and applicable to situations with 30 or fewer candidate predictors. Ongoing work aims to extend to larger numbers of candidate predictors and also to time-to-event outcomes.

For continuous outcomes, accurate predictions across the spectrum of predicted values require the standard deviation of the residuals to be precisely estimated. Supplementary material S3 shows that to target a less than 10% multiplicative error in the estimated residual standard deviation, the required sample size is simply 234+P, where P is the number of predictor parameters considered.

**Step 3: What sample size will produce a small required shrinkage of predictor effects?**

Our third recommended step is to identify the sample size required to minimise the problem of overfitting. Overfitting is when a developed model’s predictions are more extreme than they ought to be for individuals in a new dataset from the same target population. For example, an overfitted prediction model for a binary outcome will give a predicted outcome probability too close to 1 for individuals with a higher than the average outcome probability and too close to 0 for individuals with a lower than the average outcome probability. Overfitting notably occurs when the sample size is too small. In particular, when the number of candidate predictor parameters is large relative to the number of participants in total (for continuous outcomes) or to the number of participants with the outcome event (for binary or time-to-event outcomes). A consequence of overfitting is that a developed model’s apparent predictive performance (as observed in the development dataset itself) will be optimistic (i.e., too high), and its actual predictive performance in new data from the same target population will be lower (i.e., worse).

Shrinkage (also known as penalisation or regularisation) methods deal with the problem of overfitting by reducing the variability in the developed model’s predictions such that extreme predictions (e.g., predicted probabilities close to 0 or 1) are pulled back toward the overall average. However, there is no guarantee that shrinkage will fully overcome the problem of overfitting when developing a prediction model. This is because the shrinkage or penalty factors (which dictate the magnitude of shrinkage required) are also estimated from the development dataset and, especially when the sample size is small, are often imprecise and so fail to tackle the magnitude of overfitting correctly in a particular application. Furthermore, a negative correlation tends to occur between the estimated shrinkage required and the apparent performance of a model. If the apparent model performance is excellent simply by chance, the required shrinkage is typically estimated too low. Thus, ironically, in those situations when overfitting is of most concern (and thus shrinkage is most urgently needed), the prediction model developer has insufficient assurance in selecting the proper amount of shrinkage to cancel the impact of overfitting.

Riley et al therefore suggest identifying the sample size and number of candidate predictors that correspond to a small amount of desired shrinkage (≤10%) during model development. The sample size calculation (fig 3) requires the researcher to prespecify the number of candidate predictor parameters and, for binary or time-to-event outcomes, the anticipated outcome proportion or rate, respectively, in the target population. In addition, a (conservative) value for the anticipated model performance is required, as defined by the Cox-Snell R² statistic generalises to non-continuous outcomes and allows sample size calculations to minimise the expected shrinkage when developing a prediction model for binary and time-to-event outcomes (fig 3). For example, when developing a new logistic regression model with up to 20 candidate predictor parameters and an anticipated R² of 0.7, a sample size of 206 participants is required to ensure the expected shrinkage is 10% (see supplementary material S4 for full calculation). This corresponds to about seven participants for each predictor parameter considered.

The R² statistic generalises to non-continuous outcomes and allows sample size calculations to minimise the expected shrinkage when developing a prediction model for binary and time-to-event outcomes (fig 3). For example, when developing a model with a continuous outcome, true patterns are easier to detect and so overfitting is less of a concern, such that more predictor parameters can be estimated. However, when the signal:no noise ratio is low (i.e., R² is anticipated to be close to 0), true patterns are harder to identify and there is more potential for overfitting, such that fewer predictor parameters can be estimated reliably.

In the continuous outcome setting, R² is simply the coefficient of determination R², which quantifies the proportion of the variance of outcome values that is explained by the prediction model and thus is between 0 and 1. For example, when developing a prediction model for a continuous outcome with up to 30 predictor parameters and an anticipated R² of 0.7, a sample size of 206 participants is required to ensure the expected shrinkage is 10% (see supplementary material S4 for full calculation). This corresponds to about seven participants for each predictor parameter considered.

The R² statistic generalises to non-continuous outcomes and allows sample size calculations to minimise the expected shrinkage when developing a prediction model for binary and time-to-event outcomes (fig 3). For example, when developing a new logistic regression model with up to 20 candidate predictor parameters and an anticipated R² of at least 0.1, a sample size of 1698 participants is required to ensure the expected shrinkage is 10% (see fig 3 for full calculation). If the target setting has an outcome proportion of 0.3, this corresponds to an EPP of 25.5. The required sample size and EPP are sensitive to the choice of R², with lower anticipated values of R² leading to higher required sample sizes. Therefore, a conservative choice of R² is recommended (fig 4).

As in sample size calculations for randomised trials evaluating intervention effects, external evidence and expert opinion are required to inform the values that need specifying in the sample size calculator. Figure 4 provides guidance for specifying R². Importantly, unlike for continuous outcomes when R² is bounded between 0 and 1, the R² is bounded between 0 and max(R²) for binary and time-to-event outcomes. The max(R²) denotes the maximum possible value of R², which is dictated by the overall outcome proportion or rate in the development dataset and is often much less than 1. Supplementary material S5 shows the calculation of max(R²). For logistic regression models with outcome proportions of 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, and 0.01, the corresponding max(R²) values are 0.75, 0.54, 0.71, 0.63, 0.48, 0.33, and 0.11, respectively. Thus the anticipated R² might be small, even for a model with potentially good performance.
Step 4: What sample size will produce a small optimism in apparent model fit?

The sample size should also ensure a small difference in the developed models apparent and optimism adjusted values of $R^2_{\text{Nagelkerke}}$ (ie, $R^2_{\text{cs}}/\max(R^2_{\text{cs}})$), as this is a fundamental overall measure of model fit. The apparent $R^2_{\text{Nagelkerke}}$ value is simply the model's observed performance in the same data as used to develop the model, whereas the optimism adjusted $R^2_{\text{Nagelkerke}}$ value is a more realistic (approximately unbiased) estimate of the model's fit in the target population. The sample size calculations are shown in supplementary material S6 for continuous outcomes and in figure 5 for binary and time-to-event outcomes. As before, they require the user to specify the anticipated $R^2_{\text{cs}}$ and the max($R^2_{\text{cs}}$), as described in figure 4. For example, when developing a logistic regression model with an anticipated $R^2_{\text{cs}}$ of 0.2, and in a setting with an outcome proportion of 0.05 (such that the max($R^2_{\text{cs}}$)) is 0.33, 1079 participants are required to ensure the expected optimism in the apparent $R^2_{\text{Nagelkerke}}$ is just 0.05 (see figure 5 for calculation).

**Recommendations and software**

Box 1 summarises our recommended steps for calculating the minimum sample size required for prediction model development. This involves four calculations for binary outcomes (B1 to B4), three for time-to-event outcomes (T1 to T3), and four for continuous outcomes (C1 to C4). To implement the calculations, we have written the pmsampsize package for Stata and R. The software calculates the sample size needed to meet all the criteria listed in box 1 (except B2, which is available at https://mvansmeden.shinyapps.io/BeyondEPV/), conditional on the user inputting values of required parameters such as the number of candidate predictors, the anticipated outcome proportion in the target population, and the anticipated $R^2_{\text{cs}}$. The calculations are especially helpful when prospective data collection (eg, new cohort study) are required before model development; however, they are also relevant when existing data are available to guide the number of predictors that can be considered.

**Box 1 Recommendations for calculating the sample size needed when developing a clinical prediction model for continuous, binary, and time-to-event outcomes**

To increase the potential for developing a robust prediction model, the sample size should be at least large enough to minimise model overfitting and to target sufficiently precise model predictions

**Binary outcomes**

For binary outcomes, ensure the sample size is enough to:

- Estimate the overall outcome proportion with sufficient precision at one or more key time-points in follow-up (use equation in figure 1) (B1)
- Target a small mean absolute prediction error (use equation in figure 2) if number of predictor parameters is ≤30 (B2)
- Target a shrinkage factor of 0.9 (use equation in figure 3) (B3)
- Target small optimism of 0.05 in the apparent $R^2_{\text{Nagelkerke}}$ (use equation in figure 5) (B4)

**Time-to-event outcomes**

For time-to-event outcomes, ensure the sample size is enough to:

- Estimate the overall outcome proportion with sufficient precision at one or more key time-points in follow-up (use equation in figure 1) (T1)
- Target a shrinkage factor of 0.9 (use equation in figure 3) (T2)
- Target small optimism of 0.05 in the apparent $R^2_{\text{Nagelkerke}}$ (use equation in figure 5) (T3)

**Continuous outcomes**

For continuous outcomes, ensure the sample size is enough to:

- Estimate the model intercept precisely (see supplementary material 1) (C1)
- Estimate the model residual variance with sufficient precision (see supplementary material 2) (C2)
- Target a shrinkage factor of 0.9 (use equation in figure 3) (C3)
- Target small optimism of 0.05 in the apparent $R^2_{\text{Nagelkerke}}$ (use equation in figure 5) (C4)

These approaches require researchers to specify the anticipated overall outcome risk or mean outcome value in the target population, the number of candidate predictor parameters, and the anticipated model performance in terms of overall model fit ($R^2_{\text{cs}}$). When the choice of values is uncertain, we generally recommend being conservative and so taking those values (eg, smallest $R^2_{\text{cs}}$) that give larger sample sizes.

When an existing dataset is already available (such that sample size is already defined), the calculations can be used to identify if the sample size is sufficient to estimate the overall outcome risk or the mean outcome value, and how many predictor parameters can be considered before overfitting becomes a concern.

**Applied examples**

We now illustrate the recommendations in box 1 by using three examples.

**Example 1: Binary outcome**

North et al developed a model predicting pre-eclampsia in pregnant women based on clinical predictors measured at 15 weeks’ gestation, including vaginal bleeding, age, previous miscarriage, family history, smoking, and alcohol consumption. The model included 13 predictor parameters and had a C statistic of 0.71. Emerging research aims to improve this and other pre-eclampsia prediction models by including additional predictors (eg, biomarkers and ultrasound measurements).

As the outcome is binary, the sample size calculation for a new prediction model needs to examine criteria B1 to B4 in box 1. This requires us to input the overall proportion of women who will develop pre-eclampsia (0.05) and the number of candidate predictor parameters (assumed to be 30 for illustration). For an outcome proportion of 0.05, the max($R^2_{\text{cs}}$) value is 0.33 (see
supplementary material S5). If we assume, conservatively, that the new model will explain 15% of the variability, the anticipated $R^2$ value is 0.15×0.33=0.05. Now we can check criteria B1, B3, and B4 by typing in Stata:

```
psmsampsize, type(b) rsquared(0.05) parameters(30)
```

This indicates that at least 5249 women are required, corresponding to 263 events and an EPP of 8.75. This is driven by criterion B3, to ensure the expected shrinkage required is just 10% (to minimise the potential overfitting). To check criterion B2 in box 1, we can apply the formula in figure 2. This suggests that 544 women are needed to target a mean absolute error in predicted probabilities of $\pm 0.05$. This is much lower than the 5249 women needed to meet criterion B3.

If recruiting 5249 women is impractical (eg, because of time, cost, or practical constraints for data collection), the sample size required can be reduced by identifying a smaller number of candidate predictors (eg, based on existing evidence from systematic reviews4). For example, with 20 rather than 30 candidate predictors, the required sample size to meet all four criteria is at least 3500 women and 175 events (still 8.75 EPP).

**Example 2: Time-to-event outcome**

Many prognostic models are available for the risk of a recurrent venous thromboembolism (VTE) after cessation of treatment for a first VTE.12 For example, the model of Ensor et al included predictors of age, sex, site of first clot, D-dimer level, and the lag time from cessation of treatment until measurement of D-dimer (often around 30 days).13 The model's C statistic was 0.69 and the adjusted $R^2$ was 0.051 (corresponding to 8% of the total variation). Emerging research aims to extend such models by including additional predictors.

The sample size required for a new model must at least meet criteria T1 to T3.15 This requires us to input a key time point for prediction of VTE recurrence risk (eg, two years), and the conservative value of $R^2$ = 0.15×0.33=0.05. Now criteria T1 to T3 can be checked, for example by typing in Stata:

```
psmsampsize, type(s) rsquared(0.051) parameters(30) rate(0.065) timepoint(2) meanfup(2.07)
```

This indicates that at least 5143 participants are required, corresponding to 692 events and an EPP of 23.1. This is considerably more than 10 EPP, and is driven by a desired shrinkage factor (criterion T2) of only 10% to minimise overfitting based on just 8% of variation explained by the model. If the number of candidate predictor parameters is lowered to 20, the required sample size is reduced to 3429 (still an EPP of 23.1).

**Example 3: Continuous outcome**

Hudda et al developed a prediction model for fat free mass in children and adolescents aged 4 to 15 years, including 10 predictor parameters based on height, weight, age, sex, and ethnicity.16 The model is needed to provide an estimate of an individual’s current fat mass (weight minus predicted fat free mass). On external validation, the model had an $R^2$ of 0.90. Let us assume that the model will need updating (eg, in 10 years owing to changes in the population behaviour and environment), and that an additional 10 predictor parameters (and thus a total of 20 parameters) will need to be considered in the model development.

The sample size for a model development dataset must at least meet the four criteria of C1 to C4 in box 1. This requires us to specify the anticipated $R^2$ (0.90), number of candidate predictor parameters ($n$=20), and mean (26.7 kg) and standard deviation (8.7 kg) of fat free mass in the target population (taken from Hudda et al16). For example, in Stata, after installation of `psmsampsize` (type: `ssc install psmsampsize`), we can type: `psmsampsize, type(s) rsquared(0.9) parameters(20) intercept(26.7) sd(8.7)`

This returns that at least 254 participants are required, and so 12.7 participants for each predictor parameter. The sample size of 254 is driven by the number needed to precisely estimate the model standard deviation (criterion C3), as only 68 participants are needed to minimise overfitting (criteria C1 and C2).

**Extensions and further topics**

**Ensuring accurate predictions in key subgroups**

Alongside the criteria outlined in box 1, a more stringent task is to ensure model predictions are accurate in key subgroups defined by particular values or categories of included predictors.4 One way to tackle this is to ensure predictor effects in the model equation are precisely estimated, at least for key subgroups of interest.1214 For binary and time-to-event outcomes, the precision of a predictor’s effect depends on its magnitude, the variance of the predictor’s values, the predictor’s correlation with other predictors in the model, the sample size, and the outcome proportion or rate in the study.4041 For continuous outcomes, it depends on the sample size, the residual variance, the correlation of the predictor with other included predictors, and the variance of the predictor’s values.4243 Note that for important categorical predictors large sample sizes might be needed to avoid separation issues (ie, where no events or non-events occur in some categories),17 and potential bias from sparse events.18

**Sample size considerations when using an existing dataset**

Our proposed sample size calculations (ie, based on the criteria in box 1) are still useful in situations when an existing dataset is already available, with a specific number of participants and predictors. Firstly, the calculations might identify that the dataset is too small (for example, if the overall outcome risk cannot be estimated precisely) and so the collection of further data is required.1218 Secondly, the calculations might help identify how many predictors can be considered before overfitting becomes a concern. The shrinkage estimate obtained from fitting the full model (including all predictors) can be used to gauge whether the number of predictors could be reduced through data reduction techniques such as principal components analysis.10 This process should be done blind to the estimated predictor effects in the full model, as otherwise decisions about predictor inclusion will be influenced by a "quick look" at the results (which increases the overfitting).

**Sample size requirements when using variable selection**

Further research on sample size requirements with variable selection is required, especially for the use of more modern penalisation methods such as the lasso (least absolute shrinkage and selection operator) or elastic net.1019 Such methods allow shrinkage and variable selection to operate simultaneously, and they even allow the consideration of more predictor parameters.
than number of participants or outcome events (ie, in high dimensional settings). However, there is no guarantee such models solve the problem of overfitting in the dataset at hand. As mentioned, they require penalty and shrinkage factors to be estimated using the development dataset, and such estimates will often be hugely imprecise. Also, the subset of included predictors might be highly unstable64-65; that is, if the prediction model development was repeated on a different sample of the same size, a different subset of predictors might be selected and important predictors missed (especially if sample size is small). In healthcare the final set of predictors is a crucial consideration, owing to their cost, time, burden (eg, blood test, invasiveness), and measurement requirements.

**Larger sample sizes might be needed when using machine learning approaches to develop risk prediction models**

An alternative to regression based prediction models are those based on machine learning methods, such as random forests and neural networks (of which “deep learning” methods are a special case).66 When the focus is on individualised outcome risk prediction, it has been shown that extremely large datasets might be needed for machine learning techniques. For binary outcomes, machine learning techniques could need more than 10 times as many events for each predictor to achieve a small amount of overfitting compared with classic modelling techniques such as logistic regression, and might show instability and a high potential for overfitting compared with classic modelling techniques such as logistic regression, and might show instability and a high optimism even with more than 200 EPP.67 A major cause of this problem is that the number of predictor (“feature”) parameters considered by machine learning approaches will usually far exceed that for regression, even when the same set of predictors is considered, particularly because they routinely examine multiple interaction terms and categorise continuous predictors. Therefore, machine learning methods are not immune to sample size requirements, and actually might need truly big data.

**Conclusion**

Patients and healthcare professionals require clinical prediction models to accurately guide healthcare decisions.1 Larger sample sizes lead to more robust models being developed, and our guidance in box 1 outlines how to calculate the minimum sample size required. Clearly, the more data for model development, the better; so if larger sample sizes are achievable than our guidance suggests, use it! Of course, any data collected should be of sufficient quality and representative of the target population and settings of application.68-70

After data collection, careful model building is required using appropriate methods.1,71,72 In particular, we do not recommend data splitting (eg, into model training and testing samples), as this is inefficient and it is better to use all the data for model development, with resampling methods (such as bootstrapping) used for internal validation.73-75 Sometimes external information might be used to supplement the development dataset further.75-77

Lastly, sample size requirements when externally validating an existing prediction model require a different approach, as discussed elsewhere.75-77

Sample size for model updating

When an existing prediction model is updated, the existing model equation is revised using a new dataset. The required sample size for this dataset depends on how the model is to be updated and whether additional predictors are to be included. In our worked examples, we assumed that all parameters in the existing model will be re-estimated using the model updating dataset. In that situation, the researcher can still follow the guidance in box 1 for calculating the required sample size, with the total predictor parameters the same as in the original model plus those new parameters required for any additional predictors.

Sometimes, however, only a subset of the existing model’s parameters is to be updated.66,67 In particular, to deal with calibration-in-the-large, researchers might only want to revise the model intercept (or baseline survival), while constraining the other parameter estimates to be the same as those in the existing model. In this case the required sample size only needs to be large enough to estimate the mean outcome value or outcome risk precisely (ie, to meet criteria C1, B1, or T1 in box 1). Even if researchers also want to update the existing predictor effects, they might decide to constrain their updated values to be equal to the original values multiplied by a constant. Then, the sample size only needs to be large enough to estimate one predictor parameter (ie, the constant) for the existing predictors, plus any new parameters the researchers decide to add. Such model updating techniques therefore reduce the sample size needed (to meet the criteria in box 1) compared with when every predictor parameter is re-estimated without constraint.
Multivariable logistic regression in medical research: principles and practice of model building and validation

Wiley-Blackwell

1 Introduction
2 Basic principles
3 Variable selection
4 Model validation
5 Calibration
6 Other issues
7 Conclusion

References

Wiley Online Library

Abstract - Multivariable logistic regression provides a powerful technique to estimate the odds ratios of individual risk factors associated with a particular outcome. The purpose of this chapter is to introduce the principles and practice of model building and validation and to guide the reader to avoid common pitfalls in medical research. Model building involves selecting a list of variables to fit the model and fitting the model on the data. Model validation involves assessing the goodness of fit of the model and assessing the performance of the model in the population from which the data are drawn. Model accuracy is assessed by estimating the optimism of the model through cross-validation and bootstrapping. This chapter is an improved version of previous reviews and will be of interest to those who use logistic regression in medical research.

1 Introduction

1.1 Motivation

The need for better prediction models.

1.2 Aims

This chapter aims to introduce the principles and practice of multivariable logistic regression model building and validation.

1.3 Scope

This chapter is an updated version of previous reviews and will be of interest to those who use logistic regression in medical research.

2 Basic principles

2.1 Logit vs. odds

The logit (log-odds) transformation is a useful way to express the relationship between a binary outcome and a list of risk factors. It is a linear transformation of the odds ratio, and it is used to fit the logistic regression model.

2.2 Model

The logistic regression model is a linear predictor of the log-odds of the outcome, which is expressed as a linear combination of the risk factors:

logit(p) = ln(p/(1-p)) = β0 + β1x1 + β2x2 + ... + βnxn

where p is the probability of the outcome, and β0...βn are the coefficients associated with the risk factors x1...xn.

2.3 Interpretation

The coefficients β0...βn are interpreted as the log-odds ratios, which can be exponentiated to obtain the odds ratios.

exp(β0) = odds ratio for the baseline risk factor

exp(β1) = odds ratio for risk factor x1

exp(β2) = odds ratio for risk factor x2

... exp(βn) = odds ratio for risk factor xn

2.4 Multicollinearity

Multicollinearity occurs when two or more risk factors are highly correlated, which can lead to unstable estimates of the coefficients. In such cases, ridge regression or lasso regression can be used to stabilize the estimates.

2.5 Overfitting

Overfitting occurs when the model is too complex, with too many risk factors, and it is not generalizable to new populations. In such cases, cross-validation and bootstrapping can be used to assess the model's performance.

2.6 Model validation

Model validation involves assessing the goodness of fit of the model and assessing the performance of the model in the population from which the data are drawn. Model accuracy is assessed by estimating the optimism of the model through cross-validation and bootstrapping.

2.7 Discussion

This chapter is an updated version of previous reviews and will be of interest to those who use logistic regression in medical research.

3 Variable selection

3.1 Methods

Variable selection involves choosing a list of variables to fit the model and fitting the model on the data. There are several methods for variable selection, such as stepwise selection, forward selection, backward elimination, and penalized regression.

3.2 Stepwise selection

Stepwise selection involves adding or removing variables based on their statistical significance. It is a simple and intuitive method, but it can lead to overfitting and unstable estimates.

3.3 Forward selection

Forward selection involves adding variables to the model one at a time, based on their statistical significance, until no more variables improve the model's fit.

3.4 Backward elimination

Backward elimination involves removing variables from the model one at a time, based on their statistical significance, until all remaining variables improve the model's fit.

3.5 Penalized regression

Penalized regression involves adding a penalty term to the likelihood function to penalize complex models. This can reduce overfitting and improve the model's generalizability.

3.6 Model interpretation

The selected variables should be interpreted in the context of the study's research question and the clinical setting.

3.7 Discussion

This chapter is an updated version of previous reviews and will be of interest to those who use logistic regression in medical research.

4 Model validation

4.1 Model selection

Model selection involves choosing a list of variables to fit the model and fitting the model on the data. There are several methods for model selection, such as cross-validation and bootstrapping.

4.2 Cross-validation

Cross-validation involves splitting the data into training and testing sets, fitting the model on the training set, and evaluating the model's performance on the testing set. This process is repeated several times, and the average performance is used to assess the model's accuracy.

4.3 Bootstrapping

Bootstrapping involves resampling the data with replacement, fitting the model on the resampled data, and evaluating the model's performance on the original data. This process is repeated several times, and the average performance is used to assess the model's accuracy.

4.4 Discussion

This chapter is an updated version of previous reviews and will be of interest to those who use logistic regression in medical research.

5 Calibration

5.1 Interpretation

Calibration is a measure of the agreement between the predicted probabilities and the observed probabilities. A well-calibrated model produces predictions that are close to the observed outcomes.

5.2 Model calibration

Model calibration involves assessing the agreement between the predicted probabilities and the observed probabilities. This can be done using methods such as the Hosmer-Lemeshow test and the calibration slope.

5.3 Discussion

This chapter is an updated version of previous reviews and will be of interest to those who use logistic regression in medical research.

6 Other issues

6.1 Model shrinkage

Model shrinkage occurs when the coefficients are reduced in magnitude after fitting the model on the data. This can improve the model's generalizability and reduce overfitting.

6.2 Model interpretation

The selected variables should be interpreted in the context of the study's research question and the clinical setting.

6.3 Discussion

This chapter is an updated version of previous reviews and will be of interest to those who use logistic regression in medical research.

7 Conclusion

This chapter is an improved version of previous reviews and will be of interest to those who use logistic regression in medical research.
Figures

**Binary outcomes:** For a binary outcome, an approximate 95% confidence interval for the overall outcome proportion (\(\hat{\phi}\)) is,

\[
\hat{\phi} \pm 1.96 \sqrt{\frac{\hat{\phi}(1 - \hat{\phi})}{n}}
\]

and so the absolute margin of error (\(\delta\)) is

\[
1.96 \sqrt{\frac{\hat{\phi}(1 - \hat{\phi})}{n}}
\]

Thus, to aim for precise estimation of the overall outcome probability in the target population, based on the anticipated outcome proportion (\(\hat{\phi}\)) and the desired margin of error, the required sample size is calculated as:

\[
n = \left(\frac{1.96}{\delta}\right)^2 \frac{\hat{\phi}(1 - \hat{\phi})}{0.5(1 - 0.5)}
\]

We generally recommend aiming for a margin of error ≤0.05. Then assuming an anticipated outcome proportion in the study population of 0.5,

\[
n = \left(\frac{1.96}{0.05}\right)^2 \frac{0.5(1 - 0.5)}{0.5(1 - 0.5)} = 384.2
\]

and thus at least 385 participants (ie, about 385×0.5=193 participants with the outcome) are required to target an estimation error of at most 0.05 around the true value of 0.5. When the outcome proportion is 0.1 we require at least 139 participants, and an outcome proportion of 0.2 requires at least 246 participants.

**Time-to-event outcomes:** For time-to-event data, we must consider the precision of the estimated cumulative outcome incidence for a particular time point (\(t\)) of interest. A simple approach is to assume an exponential survival model (ie, constant outcome event rate over time),\(^1\) for which the cumulative incidence function is

\[
F(t) = 1 - \exp(-\hat{\lambda} t)
\]

where \(\hat{\lambda}\) is the estimated rate (number of outcome events per person year) and \(t\) is time, say in years. The variance of the estimated rate is \(\frac{\hat{\lambda}}{T}\), where \(T\) is the total person years of follow-up. This leads to an approximate 95% confidence interval for \(F(t)\) of:

\[
1 - \exp\left(-\left(\hat{\lambda} \pm 1.96 \sqrt{\frac{\hat{\lambda}}{T}}\right) t\right)
\]

If the anticipated outcome event rate (\(\lambda\)) and time point of key interest (\(t\)) in the target population are prespecified, we can calculate the total number of person years of follow-up, \(T\), that would help ensure a narrow confidence interval for \(F(t)\), such that the lower and upper bounds are no more than an absolute value of, say, 0.05 from the estimated \(F(t)\).

For example, if \(t=10\) years is of interest for prediction, and there is an assumed outcome event rate of 0.10 (10 events per 100 person years), then assuming the exponential model we obtain \(F(10) = 1 - \exp(-0.1 \times 10) = 0.632\). To target a margin of error in the estimate of this outcome risk of ≤0.05 we require at least 2366 person years of follow-up because,

\[
1 - \exp\left(-\left(\hat{\lambda} \pm 1.96 \sqrt{\frac{\hat{\lambda}}{T}}\right) t\right) = 1 - \exp\left(-\left(0.10 \pm 1.96 \sqrt{\frac{0.10}{2366}}\right) 10\right) = 0.582\text{ to }0.676
\]

and so the 95% confidence interval has lower and upper bounds ≤0.05 of the true value of 0.632.

**Fig 1** Calculation of sample size required for precise estimation of the overall outcome probability in the target population.
For a binary outcome van Smeden et al. use simulation, across a range of scenarios, to derive an approximation (on the natural log scale, denoted by ln) of the expected average error in the outcome probabilities when a derived model is applied to new individuals from the target population. Their derived formula was originally developed based on 12 or fewer predictor parameters, but we have since updated the simulations to allow for 30 or fewer predictor parameters. The derived formula is:

\[ \ln(\text{MAPE}) = -0.508 - 0.544 \ln(n) + 0.259 \ln(\Phi) + 0.504 \ln(P) \]

Here, \( n \) is the sample size of the development dataset, \( \Phi \) is the anticipated outcome proportion (≤0.5), and \( P \) is the number of candidate predictor parameters (≤30). MAPE denotes the Mean Absolute Prediction Error (i.e., the average error in the model’s estimated outcome probability one would allow for in the intended setting of application of the model). Rearranging this equation, and choosing a target value for MAPE, the required sample size is:

\[ n = \exp \left( \frac{-0.508 + 0.259 \ln(\Phi) + 0.504 \ln(P) - \ln(\text{MAPE})}{0.544} \right) \]

We recommend that MAPE is no larger than 0.050, but lower values might be appropriate in settings where precise predictions are demanded if the consequences of wrong decisions are large. For example, if we set MAPE to 0.050, in a setting with an anticipated outcome proportion 0.30 and 10 candidate predictor parameters, we require

\[ n = \exp \left( \frac{-0.508 + 0.259 \ln(0.30) + 0.504 \ln(10) - \ln(0.050)}{0.544} \right) = 460.9 \]

and thus at least 461 participants (about 138 events) in the development dataset, corresponding to an EPP of 13.8. If MAPE is reduced to 0.04, the development dataset requires at least 695 participants (about 209 expected events and an EPP of 20.8). In this sample size equation \( \Phi \) is the outcome proportion if it is ≤0.5. If the outcome proportion is >0.5, then researchers should rather specify \( \Phi = 1 - \) outcome proportion.

**Fig 2** Sample size required to help ensure a developed prediction model of a binary outcome will have a small mean absolute error in predicted probabilities when applied in other targeted individuals.

For binary or time-to-event outcomes, Riley et al. show that the sample size (number of participants, \( n \)) needed to achieve an expected uniform shrinkage factor of \( S \) can be expressed as:

\[ n = \frac{P}{(S - 1) \ln \left( 1 - \frac{R^2_{\text{CS}}}{S} \right)} \]

We suggest targeting a shrinkage of ≤10%, such that \( S ≥ 0.9 \). For example, for developing a new logistic regression model based on up to 20 candidate predictor parameters with an anticipated \( R^2_{\text{CS}} \) of at least 0.1, then to target an expected shrinkage of 0.9 we need a sample size of:

\[ n = \frac{P}{(S - 1) \ln \left( 1 - \frac{R^2_{\text{CS}}}{S} \right)} = \frac{20}{(0.9 - 1) \ln \left( 1 - \frac{0.1}{0.9} \right)} = 1698 \]

and thus 1698 participants. If the target population has an outcome proportion of 0.1, then the 1698 participants corresponds to 1698 x 0.1 = 169.8 outcome events. With 20 predictor parameters, the required events per candidate predictor parameter (EPP) = (1698 x 0.1)/20 = 8.5. However, if the target setting has an outcome proportion of 0.3, the EPP is 25.5. The big change in the required EPP is because, although the chosen value of \( R^2_{\text{CS}} \) is fixed at 0.1, the maximum value of \( R^2_{\text{CS}} \) is much higher for the setting with the higher outcome proportion (see supplementary material S1).

**Fig 3** How to calculate the sample size needed to target a small magnitude of required shrinkage of predictor effects (to minimise potential model overfitting) for binary or time-to-event outcomes.
The sample size equations in figure 3 require a (conservative) value for the model’s anticipated $R^2_{CS}$ (proportion of overall variation explained) to be prespecified. A sensible value for this can be obtained in various ways.

**Using values reported directly for existing models**
Sometimes $R^2_{CS}$ is reported directly or can be requested for previous prediction model studies for the same (or similar) target population, considering the same (or similar) outcomes and (if relevant) time points of interest. For example, the prediction model developer could consult systematic reviews of similar prediction models or registries that record existing prediction models available in a particular clinical topic area. Indeed, often a new prediction model is developed specifically to update or improve (eg, by adding additional predictors) on the performance of an existing model that was developed in the same setting and target population (with similar outcome proportion) of interest. Then, this existing model’s $R^2_{CS}$ could be used as a conservative value for the new model’s anticipated $R^2_{CS}$.

**Using values based on other performance measures reported for an existing model**
For situations when $R^2_{CS}$ is not directly reported for an existing model, Riley et al show how it can be derived from other reported information, such as the likelihood ratio statistic, the C statistic (area under the curve), Royston’s D statistic, and other pseudo $R^2$ measures, such as Nagelkerke’s $R^2$, McFadden’s $R^2$, O’Quigley’s $R^2$, Royston’s $R^2$, and Royston and Sauerbrei’s $R^2$. When $R^2_{CS}$ is extracted from a model development study, ideally it should be adjusted for optimism due to any overfitting, to give a more honest (unbiased) estimate of performance.

**Deciding values in the absence of existing information**
In the absence of any other information, we suggest that sample sizes be derived assuming the value of $R^2_{CS}$ corresponds to an $R^2_{Nagelkerke}$ of 0.15 (ie, $R^2_{CS} = 0.15 \times \max (R^2_{CS})$), such that 15% of the total variance is explained. The max ($R^2_{CS}$) is 1 for continuous outcomes, but usually less than 1 for binary and time-to-event outcomes (see explanation in supplementary material S5). Medical diagnosis and prediction of health related outcomes are, generally speaking, low signal/noise ratio situations, and it is not uncommon to see $R^2_{Nagelkerke}$ values in the 0.1 to 0.2 range. An exception is when predictors include “direct” (mechanistic) measurements, such as including the baseline version of the binary or ordinal outcome (eg, including smoking status at baseline when predicting smoking status at one year), or direct measures of the processes involved (eg, including physiological function of patients in intensive care when predicting risk of death within 48 hours). Then, in this special situation, an $R^2_{CS} = 0.5$ may be a more appropriate default choice, such that $R^2_{CS} = 0.5 \times \max (R^2_{CS})$.

Fig 4 How to decide on the model’s anticipated $R^2_{CS}$ in advance of data collection
The aim is to calculate the sample size required to ensure a small expected optimism in the apparent $R^2_{\text{Nagelkerke}}$ (i.e. $R^2_{\text{CS}}/\max(R^2_{\text{CS}})$). For binary or time-to-event outcomes, this firstly requires the calculation of the shrinkage factor that corresponds to an expected optimism of $\delta$ in $R^2_{\text{Nagelkerke}}$. The solution provided by Riley et al.\textsuperscript{15} is:

$$S = \frac{R^2_{\text{CS}}}{R^2_{\text{CS}} + \delta \max(R^2_{\text{CS}})}$$

We suggest $\delta$ is a small value, such as $\leq 0.05$. The obtained value of $S$ can then be placed into the equation derived for binary and time-to-event outcomes in figure 3; that is:

$$n = \frac{p}{(S - 1) \ln \left(1 - \frac{R^2_{\text{CS}}}{S}\right)}$$

For example, consider the development of a logistic regression model with anticipated $R^2_{\text{CS}}$ of at least 0.2, and in a setting with an outcome proportion of 0.05, such that the $\max(R^2_{\text{CS}})$ is 0.33 (as explained in supplementary material S3). Then, to aim for an $\delta$ of $\leq 0.05$, we require:

$$S = \frac{0.2}{0.2 + (0.05 \times 0.33)} = 0.924$$

If there will be 20 candidate predictor parameters, this leads to a required sample size of:

$$n = \frac{20}{(0.924 - 1) \ln \left(1 - \frac{0.2}{0.924}\right)} = 1078.9$$

Hence, a total of at least 1079 participants are required to ensure a small expected optimism of $\leq 0.05$ in the apparent $R^2_{\text{Nagelkerke}}$.

Fig 5 How to calculate the sample size needed to target a small optimism in model fit (to minimise potential model overfitting) for binary and time-to-event outcomes