thebmj

BMJ 2020;368:m441 doi: 10.1136/bmj.m441 (Published 18 March 2020)

RESEARCH METHODS & REPORTING

Check for updates

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 biostatistics*[1], Joie Ensor *lecturer in biostatistics*[1], Kym I E Snell *lecturer in biostatistics*[1], Frank E Harrell Jr *professor of biostatistics*[2], Glen P Martin *lecturer in health data sciences*[3], Johannes B Reitsma *associate professor*[4], Karel G M Moons *professor of clinical epidemiology*[4], Gary Collins *professor of medical statistics*[5], Maarten van Smeden *assistant professor*[4 5 6]

[1]Centre for Prognosis Research, School of Primary, Community and Social Care, Keele University, Staffordshire ST5 5BG, UK; [2]Department of Biostatistics, Vanderbilt University School of Medicine, Nashville TN, USA; [3]Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK; [4]Julius Center for Health Sciences, University Medical Center Utrecht, Utrecht, Netherlands; [5]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK; [6]Department 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.[1-3] Well known examples include the Wells score,[4 5] QRISK,[6 7] and the Nottingham prognostic index.[8 9] 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).[10] 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 prediction[11]) and observed outcome.

Correspondence to: R D Riley r.riley@keele.ac.uk (or @richard_d_riley on Twitter)

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,[12] with various blanket "rules of thumb" proposed and debated.[13-17] 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.[13-16 18] 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 about 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.[10] 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.[19-21] This is widely referred to as needing at least 10 events per variable (10 EPV). 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,[19-21] 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.[10 22-26] 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.[13-17] 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,[27] but these share the same concerns as for 10 EPP.[16]

Sample size calculation to ensure precise predictions and minimise overfitting

Recent work by van Smeden et al[13 14] and Riley et al[15 16] 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 (R^2). 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).[15] 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).[16]

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.[14] 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.[10 16] 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.[28] 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 (ie, too high), and its actual predictive performance in new data from the same target population will be lower (ie, 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 (eg, predicted probabilities close to 0 or 1) are pulled back toward the overall average.[29-34] 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.[30] 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.[30] 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.[15 16] 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 R squared statistic (R^2_{cs}).[15 35] The anticipated value of R^2_{cs} is important because it reflects the signal:noise ratio, which has an impact on the estimation of multiple parameters and the potential for overfitting. When the signal:noise ratio is anticipated to be high (eg, R^2_{cs} is close to 1 for a prediction 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:noise ratio is low (ie, R^2_{cs} 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, R^2_{cs} is simply the coefficient of determination R^2, 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 R^2_{cs} 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 R^2_{cs} 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 R^2_{cs} 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 R^2_{cs}, with lower anticipated values of R^2_{cs} leading to higher required sample sizes. Therefore, a conservative choice of R^2_{cs} 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 R^2_{cs}. Importantly, unlike for continuous outcomes when R^2_{cs} is bounded between 0 and 1, the R^2_{cs} is bounded between 0 and $\max(R^2_{cs})$ for binary and time-to-event outcomes. The $\max(R^2_{cs})$ denotes the maximum possible value of R^2_{cs}, 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 $\max(R^2_{cs})$. 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 $\max(R^2_{cs})$ values are 0.75, 0.74, 0.71, 0.63, 0.48, 0.33, and 0.11, respectively. Thus the anticipated R^2_{cs} 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 $R^2_{Nagelkerke}$ (ie, $R^2_{cs}/max(R^2_{cs})$), as this is a fundamental overall measure of model fit.[10 38] The apparent $R^2_{Nagelkerke}$ value is simply the model's observed performance in the same data as used to develop the model, whereas the optimism adjusted $R^2_{Nagelkerke}$ 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_{cs} and the $max(R^2_{cs})$, as described in figure 4. For example, when developing a logistic regression model with an anticipated R^2_{cs} of 0.2, and in a setting with an outcome proportion of 0.05 (such that the $max(R^2_{cs})$ is 0.33), 1079 participants are required to ensure the expected optimism in the apparent $R^2_{Nagelkerke}$ 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_{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 (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 $R^2_{Nagelkerke}$ (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 $R^2_{Nagelkerke}$ (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 $R^2_{Nagelkerke}$ (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_{cs}). When the choice of values is uncertain, we generally recommend being conservative and so taking those values (eg, smallest R^2_{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,[43] 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 $max(R^2_{cs})$ 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_{cs} value is 0.15×0.33=0.05. Now we can check criteria B1, B3, and B4 by typing in Stata:

```
pmsampsize, type(b) rsquared(0.05) parameters(30)
prevalence(0.05)
```

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 ≤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 reviews[44]). 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.[45] 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).[46] The model's C statistic was 0.69 and the adjusted R^2_{cs} 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), alongside the number of candidate predictor parameters (n=30), the anticipated mean follow-up (2.07 years), and outcome event rate (0.065, or 65 VTE recurrences for every 1000 person years of follow-up), and the conservative value of R^2_{cs} (0.051), with all chosen values based on Ensor et al.[46] Now criteria T1 to T3 can be checked, for example by typing in Stata:

```
pmsampsize, 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.[47] 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_{cs} 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_{cs} (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 al[47]). For example, in Stata, after installation of pmsampsize (type: ssc install pmsampsize), we can type:

```
pmsampsize, type(c) 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.[48] One way to tackle this is to ensure predictor effects in the model equation are precisely estimated, at least for key subgroups of interest.[15 16] 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.[49-51] 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.[48 52-55] 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),[13] and potential bias from sparse events.[56]

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.[57 58] 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.[33 59] 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 unstable[60-63]; 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).[64] 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 optimism even with more than 200 EPP.[26] 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" to ensure their developed models have small overfitting, and for their potential advantages (eg, dealing with highly non-linear relations and complex interactions) to reach fruition. The size of most medical research datasets is better suited to using regression (including penalisation and shrinkage approaches),[65] especially as regression also leads to a transparent model equation that facilitates implementation, validation, and graphical displays.

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.

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 69]

After data collection, careful model building is required using appropriate methods.[1 3 10] 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.[70 71] Sometimes external information might be used to supplement the development dataset further.[72-74] Lastly, sample size requirements when externally validating an existing prediction model require a different approach, as discussed elsewhere.[75-78]

Contributors: RDR and MVS led the methodology that underpins the methods in this article, with contributions from all authors. RDR wrote the first and updated drafts of the article, with important contributions and revisions from all authors at multiple stages. JE led the development of the pmsampsize packages in Stata and R. RDR is the guarantor.

Funding: KGMM receives funding from the Netherlands Organisation for Scientific Research (project 9120.8004 and 918.10.615). For his work on this paper, FEH was supported by CTSA award No UL1 TR002243 from the National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent official views of the National Center for Advancing Translational Sciences or the US National Institutes of Health. GC is supported by the National Institute for Health Research (NIHR) Biomedical Research Centre, Oxford. KIES is funded by the NIHR School for Primary Care Research. This (publication/paper/report) presents independent research funded by the NIHR. The views expressed are those of the authors and not necessarily those of the National Health Service, NIHR, or Department of Health and Social Care.

Competing interests: We have read and understood the BMJ Group policy on declaration of interests and declare we have no competing interests.

Provenance and peer review: Not commissioned; peer reviewed.

Patient and public involvement: Patients or the public were not involved in the design, conduct, reporting, or dissemination of our research.

Dissemination to participants and related patient and public communities: We plan to disseminate the sample size calculations to our Patient and Public Involvement and Engagement team when they are applied in new research projects.

The lead author (RDR) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

1 Riley RD, van der Windt D, Croft P, etal , eds. *Prognosis Research in Healthcare: Concepts, Methods and Impact.* Oxford University Press, 2019.
2 Steyerberg EW, Moons KG, van der Windt DA, etal. PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10:e1001381. 10.1371/journal.pmed.1001381 23393430
3 Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating.* Springer, 200910.1007/978-0-387-77244-8.

4 Wells PS, Anderson DR, Rodger M, etal . Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED D-dimer. *Thromb Haemost* 2000;83:416-20. 10.1055/s-0037-1613830 10744147
5 Wells PS, Anderson DR, Bormanis J, etal . Value of assessment of pretest probability of deep-vein thrombosis in clinical management. *Lancet* 1997;350:1795-8. 10.1016/S0140-6736(97)08140-3 9428249
6 Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. *Am Heart J* 1991;121:293-8. 10.1016/0002-8703(91)90861-B 1985385
7 Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017;357:j2099. 10.1136/bmj.j2099 28536104
8 Haybittle JL, Blamey RW, Elston CW, etal . A prognostic index in primary breast cancer. *Br J Cancer* 1982;45:361-6. 10.1038/bjc.1982.62 7073932
9 Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res Treat* 1992;22:207-19. 10.1007/BF01840834 1391987
10 Harrell FEJr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. Springer, 201510.1007/978-3-319-19425-7.
11 Whittle R, Royle KL, Jordan KP, Riley RD, Mallen CD, Peat G. Prognosis research ideally should measure time-varying predictors at their intended moment of use. *Diagn Progn Res* 2017;1:1. 10.1186/s41512-016-0006-6 31093533
12 Moons KG, Altman DG, Reitsma JB, etal . Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1-73. 10.7326/M14-0698 25560730
13 van Smeden M, de Groot JA, Moons KG, etal . No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol* 2016;16:163. 10.1186/s12874-016-0267-3 27881078
14 van Smeden M, Moons KG, de Groot JA, etal . Sample size for binary logistic prediction models: Beyond events per variable criteria. *Stat Methods Med Res* 2019;28:2455-74. 10.1177/962280218784726. 29966490
15 Riley RD, Snell KI, Ensor J, etal . Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* 2019;38:1276-96. 10.1002/sim.7992 30357870
16 Riley RD, Snell KIE, Ensor J, etal . Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. *Stat Med* 2019;38:1262-75. 10.1002/sim.7993 30347470
17 Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *J Clin Epidemiol* 2011;64:993-1000. 10.1016/j.jclinepi.2010.11.012 21411281
18 de Jong VMT, Eijkemans MJC, van Calster B, etal . Sample size considerations and predictive performance of multinomial logistic prediction models. *Stat Med* 2019;38:1601-19. 10.1002/sim.8063 30614028
19 Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol* 1995;48:1503-10. 10.1016/0895-4356(95)00048-8 8543964
20 Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;49:1373-9. 10.1016/S0895-4356(96)00236-3 8970487
21 Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. *J Clin Epidemiol* 1995;48:1495-501. 10.1016/0895-4356(95)00510-2 8543963
22 Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol* 2007;165:710-8. 10.1093/aje/kwk052 17182981
23 Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. *J Clin Epidemiol* 2016;76:175-82. 10.1016/j.jclinepi.2016.02.031 26964707
24 Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat Methods Med Res* 2017;26:796-808. 10.1177/0962280214558972 25411322
25 Wynants L, Bouwmeester W, Moons KG, etal . A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. *J Clin Epidemiol* 2015;68:1406-14. 10.1016/j.jclinepi.2015.02.002 25817942
26 van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014;14:137. 10.1186/1471-2288-14-137 25532820
27 Austin PC, Steyerberg EW. The number of subjects per variable required in linear regression analyses. *J Clin Epidemiol* 2015;68:627-36. 10.1016/j.jclinepi.2014.12.014 25704724
28 Harrell FEJr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361-87. 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4 8668867
29 Pavlou M, Ambler G, Seaman SR, etal . How to develop a more accurate risk prediction model when there are few events. *BMJ* 2015;351:h3868. 10.1136/bmj.h3868 26264962
30 Van Houwelingen JC. Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Stat Neerl* 2001;55:17-3410.1111/1467-9574.00154.
31 Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. *Stat Med* 1990;9:1303-25. 10.1002/sim.4780091109 2277880
32 Copas JB. Regression, Prediction and Shrinkage. *J R Stat Soc B* 1983;45:311-5410.1111/j.2517-6161.1983.tb01258.x.
33 Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B* 1996;58:267-8810.1111/j.2517-6161.1996.tb02080.x.
34 Copas JB. Using regression models for prediction: shrinkage and regression to the mean. *Stat Methods Med Res* 1997;6:167-83. 10.1177/096228029700600206 9261914
35 Cox DR, Snell EJ. *The Analysis of Binary Data*. 2nd ed. Chapman and Hall, 1989.
36 Debray TP, Damen JA, Snell KI, etal . A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017;356:i6460. 10.1136/bmj.i6460 28057641
37 Wessler BS, Paulus J, Lundquist CM, etal . Tufts PACE Clinical Predictive Model Registry: update 1990 through 2015. *Diagn Progn Res* 2017;1:20. 10.1186/s41512-017-0021-2 31093549
38 Nagelkerke N. A note on a general definition of the coefficient of determination. *Biometrika* 1991;78:691-2. 10.1093/biomet/78.3.691.
39 McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, ed. *Frontiers in Econometrics New York*. Academic Press, 1974: 104-42.
40 O'Quigley J, Xu R, Stare J. Explained randomness in proportional hazards models. *Stat Med* 2005;24:479-89. 10.1002/sim.1946 15532086
41 Royston P. Explained variation for survival models. *Stata J* 2006;6:83-9610.1177/1536867X0600600105.
42 Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;23:723-48. 10.1002/sim.1621 14981672
43 North RA, McCowan LM, Dekker GA, etal . Clinical risk prediction for pre-eclampsia in nulliparous women: development of model in international prospective cohort. *BMJ* 2011;342:d1875. 10.1136/bmj.d1875 21474517
44 Riley RD, Moons KGM, Snell KIE, etal . A guide to systematic review and meta-analysis of prognostic factor studies. *BMJ* 2019;364:k4597. 10.1136/bmj.k4597 30700442
45 Ensor J, Riley RD, Moore D, Snell KI, Bayliss S, Fitzmaurice D. Systematic review of prognostic models for recurrent venous thromboembolism (VTE) post-treatment of first unprovoked VTE. *BMJ Open* 2016;6:e011190. 10.1136/bmjopen-2016-011190 27154483
46 Ensor J, Riley RD, Jowett S, etal. PIT-STOP collaborative group. Prediction of risk of recurrence of venous thromboembolism following treatment for a first unprovoked venous thromboembolism: systematic review, prognostic model and clinical decision rule, and economic evaluation. *Health Technol Assess* 2016;20:i-xxxiii, 1-190. 10.3310/hta20120 26879848
47 Hudda MT, Fewtrell MS, Haroun D, etal . Development and validation of a prediction model for fat mass in children and adolescents: meta-analysis using individual participant data. *BMJ* 2019;366:l4293. 10.1136/bmj.l4293 31340931
48 Maxwell SE, Kelley K, Rausch JR. Sample size planning for statistical power and accuracy in parameter estimation. *Annu Rev Psychol* 2008;59:537-63. 10.1146/annurev.psych.59.103006.093735 17937603
49 Hsieh FY, Lavori PW, Cohen HJ, Feussner JR. An overview of variance inflation factors for sample-size calculation. *Eval Health Prof* 2003;26:239-57. 10.1177/0163278703255230 12971199
50 Hsieh FY, Lavori PW. Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Control Clin Trials* 2000;21:552-60. 10.1016/S0197-2456(00)00104-5 11146149
51 Schmoor C, Sauerbrei W, Schumacher M. Sample size considerations for the evaluation of prognostic factors in survival analysis. *Stat Med* 2000;19:441-52. 10.1002/(SICI)1097-0258(20000229)19:4<441::AID-SIM349>3.0.CO;2-N 10694729
52 McClelland GH. Increasing statistical power without increasing sample size. *Am Psychol* 2000;55:963-410.1037/0003-066X.55.8.963.
53 Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. *Stat Med* 1998;17:1623-34. 10.1002/(SICI)1097-0258(19980730)17:14<1623::AID-SIM871>3.0.CO;2-S 9699234
54 Kelley K, Maxwell SE. Sample size for multiple regression: obtaining regression coefficients that are accurate, not simply significant. *Psychol Methods* 2003;8:305-21. 10.1037/1082-989X.8.3.305 14596493
55 Kelley K. Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. *Behav Res Methods* 2007;39:755-66. 10.3758/BF03192966 18183888
56 Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *BMJ* 2016;352:i1981. 10.1136/bmj.i1981 27121591
57 Riley RD, Ensor J, Snell KI, etal . External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016;353:i3140. 10.1136/bmj.i3140 27334381
58 Debray TPA, Riley RD, Rovers MM, Reitsma JB, Moons KGCochrane IPD Meta-analysis Methods group. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. *PLoS Med* 2015;12:e1001886. 10.1371/journal.pmed.1001886 26461078
59 Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. *J R Stat Soc Series B Stat Methodol* 2005;67:301-2010.1111/j.1467-9868.2005.00503.x.
60 Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. *Biom J* 2018;60:431-49. 10.1002/bimj.201700067 29292533
61 Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. *Stat Med* 1992;11:2093-109. 10.1002/sim.4780111607 1293671
62 Sauerbrei W, Boulesteix AL, Binder H. Stability investigations of multivariable regression models derived from low- and high-dimensional data. *J Biopharm Stat* 2011;21:1206-31. 10.1080/10543406.2011.629890 22023687
63 Royston P, Sauerbrei W. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Stat Med* 2003;22:639-59. 10.1002/sim.1310 12590419
64 Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer, 200910.1007/978-0-387-84858-7.
65 Steyerberg EW, van der Ploeg T, Van Calster B. Risk prediction with machine learning and regression methods. *Biom J* 2014;56:601-6. 10.1002/bimj.201300297 24615859
66 Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;23:2567-86. 10.1002/sim.1844 15287085
67 Vergouwe Y, Nieboer D, Oostenbrink R, etal . A closed testing procedure to select an appropriate method for updating prediction models. *Stat Med* 2017;36:4529-39. 10.1002/sim.7179 27891652
68 Wolff RF, Moons KGM, Riley RD, etal. PROBAST Group†. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med* 2019;170:51-8. 10.7326/M18-1376 30596875
69 Moons KGM, Wolff RF, Riley RD, etal . PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann Intern Med* 2019;170:W1-33. 10.7326/M18-1377 30596876
70 Steyerberg EW, Harrell FEJJr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001;54:774-81. 10.1016/S0895-4356(01)00341-9 11470385
71 Steyerberg EW, Harrell FEJr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;69:245-7. 10.1016/j.jclinepi.2015.04.005 25981519
72 Debray TP, Koffijberg H, Lu D, Vergouwe Y, Steyerberg EW, Moons KG. Incorporating published univariable associations in diagnostic and prognostic modeling. *BMC Med Res Methodol* 2012;12:121. 10.1186/1471-2288-12-121 22883206

73 Debray TP, Koffijberg H, Nieboer D, Vergouwe Y, Steyerberg EW, Moons KG. Meta-analysis and aggregation of multiple published prediction models. *Stat Med* 2014;33:2341-62. 10.1002/sim.6080 24752993

74 Debray TP, Koffijberg H, Vergouwe Y, Moons KG, Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. *Stat Med* 2012;31:2697-712. 10.1002/sim.5412 22733546

75 Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med* 2016;35:214-26. 10.1002/sim.6787 26553135

76 Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016;74:167-76. 10.1016/j.jclinepi.2015.12.005 26772608

77 Jinks RC, Royston P, Parmar MK. Discrimination-based sample size calculations for multivariable prognostic models for time-to-event data. *BMC Med Res Methodol* 2015;15:82. 10.1186/s12874-015-0078-y 26459415

78 Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005;58:475-83. 10.1016/j.jclinepi.2004.06.017 15845334

Accepted: 19 12 2019

Figures

Binary outcomes: For a binary outcome, an approximate 95% confidence interval for the overall outcome proportion (ϕ) is,

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

and so the absolute margin of error (δ) 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:

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

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

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

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),[15] 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 (λ) 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[14] 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, Φ is the anticipated outcome proportion (≤0.5), and P is the number of candidate predictor parameters (≤30). MAPE denotes the Mean Absolute Prediction Error (ie, 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 when 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 ϕ is the outcome proportion if it is ≤0.5. If the outcome proportion is >0.5, then researchers should rather specify ϕ = 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[15 31] show that the sample size (number of participants, n) needed to achieve an expected uniform shrinkage factor of S can be expressed as:

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

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_{CS} of at least 0.1, then to target an expected shrinkage of 0.9 we need a sample size of,

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

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_{CS} is fixed at 0.1, the maximum value of R^2_{CS} is much higher for the setting with the higher outcome proportion (see supplementary material S5).

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[36] or registries that record existing prediction models available in a particular clinical topic area.[37] 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[15] 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,[38] McFadden's R^2,[39] O'Quigley's R^2,[40] Royston's R^2,[41] and Royston and Sauerbrei's R^2.[42] 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.[15]

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 $R^2_{Nagelkerke} = 0.5$ 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 $R^2_{Nagelkerke}$ (ie, $R^2_{CS}/\max(R^2_{CS})$). For binary or time-to-event outcomes, this firstly requires the calculation of the shrinkage factor that corresponds to an expected optimism of δ in $R^2_{Nagelkerke}$. The solution provided by Riley et al[15] is:

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

We suggest δ is a small value, such as ≤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:

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

For example, consider the development of a logistic regression model with anticipated R^2_{CS} of at least 0.2, and in a setting with an outcome proportion of 0.05, such that the $\max(R^2_{CS})$ is 0.33 (as explained in supplementary material S5). Then, to aim for an δ of ≤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 ≤0.05 in the apparent $R^2_{Nagelkerke}$.

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