Minimum sample size for developing a multivariable prediction model: Part I – continuous outcomes

Richard D Riley[1*], Kym IE Snell[1], Joie Ensor[1],

Danielle L Burke[1], Frank E Harrell Jr[2], Karel GM Moons[3], Gary S Collins[4]

Contact details:

* corresponding author:

Professor of Biostatistics; e- mail: r.riley@keele.ac.uk; tel: +44 (0) 1782 733905

[1] Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, UK. ST5 5BG

[2] Department of Biostatistics, Vanderbilt University School of Medicine, Nashville TN, USA.

[3] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

[4] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK. OX3 7LD.

Running title: Minimum sample size for prediction models with a continuous outcome

Key words: Multivariable prediction model, continuous outcome, linear regression, R-squared, minimum sample size

Word count: 6500

Abstract

In the medical literature, hundreds of prediction models are being developed to predict health outcomes in individuals. For continuous outcomes, typically a linear regression model is developed to predict an individual's outcome value conditional on values of multiple predictors (covariates). To improve model development and reduce the potential for overfitting, a suitable sample size is required in terms of the number of subjects (n) relative to the number of predictor parameters (p) for potential inclusion. We propose that the minimum value of n should meet four key criteria: (i) small optimism in predictor effect estimates as defined by a global shrinkage factor of ≥ 0.9; (ii) small absolute difference of ≤ 0.05 in the apparent and adjusted R^2; (iii) precise estimation (a margin of error $\leq 10\%$ of the true value) of the model's residual standard deviation, and similarly (iv) precise estimation of the mean predicted outcome value (model intercept). The criteria require pre-specification of the user's chosen p and the model's anticipated R^2 as informed by previous studies. The value of n that meets all four criterion provides the minimum sample size required for model development. In an applied example, a new model to predict lung function in African-American women using 25 predictor parameters requires at least 918 subjects to meet all criteria, corresponding to at least 36.7 subjects per predictor parameter. Even larger sample sizes may be needed to additionally ensure precise estimates of key predictor effects, especially when important categorical predictors have low prevalence in certain categories.

1 Introduction

Each year in the medical literature, hundreds of prediction models are developed to predict health outcomes in individuals.[1-3] Such models estimate an individual's predicted risk (for binary or categorical outcomes) or their expected outcome value (for a continuous outcome), conditional on the individual's observed value of multiple predictors. In this article, we focus on multivariable prediction models for continuous outcomes (e.g. blood pressure, birth weight, depression score), which are typically developed using linear regression. This provides an equation containing an intercept term and multiple predictor effects (corresponding to mean differences), which is then used in new individuals to predict their expected outcome value. Predictors (also known as variables, covariates or prognostic factors[4]) typically include standard characteristics, such as age and stage of disease, or increasingly, biomarkers and genetic information.

Prediction models for continuous outcomes can potentially inform healthcare decisions and patient management, for example to help decide on treatment and monitoring strategies.[1] Therefore, when developing their model, researchers should strive to use high quality datasets that allow a reliable model to be produced. This includes ensuring the dataset has a suitable sample size. In particular, the number of subjects should be large enough relative to the number of predictor parameters to be estimated; otherwise overfitting may be a serious problem. Overfitting refers to when a model is capturing idiosyncrasies in the development data; this leads to optimism in predictive performance such that the apparent performance is too high for the underlying population from which the development sample is drawn.[5] For example, in the development dataset the developed model's apparent proportion of variation explained (R^2) will often be too high, and the model's predicted outcome values will often be too extreme (i.e. pushed too far from the mean). Therefore, it is good practice to ensure sample sizes are large enough to minimise this problem.[3]

In this article, we build on previous work by Harrell,[3 6] to propose how to calculate a suitable sample size for development of a prediction model using linear regression. Specifically, we suggest that the minimum sample size required should minimise the potential for overfitting (and therefore optimism) and ensure precise estimates of key model parameters. We propose four criteria: (i) small optimism in predictor effect estimates; (ii) small absolute difference in the apparent and adjusted R^2; (iii) precise estimation of the residual standard deviation; and (iv) precise estimation of the mean predicted outcome value (model intercept when predictors are mean-centred). The number of subjects that meets all four criteria provides the minimum sample size required for model development.

The paper outline is as follows. In Section 2 we provide formulae to calculate the sample size required to meet criterion (i) and (ii) conditional on the user pre-specifying the number of predictor parameters (p) and the model's anticipated proportion of variation explained (R^2), as informed by previous studies. Criteria (iii) and (iv) are then described in Section 3, and we show how to calculate sample sizes that ensure a small margin of error in the estimates, such as within 10% of their true values. Section 4 then provides an example to illustrate the approach. Section 5 briefly mentions that additional criteria may be important, such as precise estimation of predictor effects, and Section 6 concludes with discussion. A subsequent paper extends the ideas to binary and time-to-event models, such as logistic and Cox regression.[7]

2   Sample size required to minimise overfitting and optimism

To adjust for overfitting during model development, statistical methods for penalisation of predictor effect estimates are available, where regression coefficients are shrunk toward zero from their usual estimated value (e.g. from traditional maximum likelihood estimation). There are many options for shrinkage,[8] including a global shrinkage factor (sometimes referred to as a uniform shrinkage factor) that is derived and applied post-estimation,[9 10] or more holistic options such as ridge regression, elastic net and the Lasso, which operate during the estimation process.[11 12] However, the penalisation factors used within these methods are often estimated with large uncertainty, which increases as the magnitude of overfitting increases. Van Houwelingen notes that: *"… shrinkage works on the average but may fail in the particular unique problem on which the statistician is working."* [8] Therefore, it is important to minimise the potential for overfitting. In this section, we outline how researchers can target a sample size (n) to minimise the potential for overfitting in advance of model development. Our formula is motivated by the concept of a global shrinkage factor, and so we begin by introducing this.

2.1   Global shrinkage factor

Consider a continuous outcome (Y_i) for i = 1 to n subjects (participants) in a study, to which we want to fit a linear regression model of the form:

$$Y_i = \mu_i + e_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + e_i \qquad (1)$$

$$e_i \sim N(0, \sigma^2)$$

Assume that the unknown parameters of the equation (i.e. the βs and σ^2) are estimated using the data, usually via ordinary least squares or maximum likelihood estimation. The intercept term, α, is

the true mean outcome value for individuals whose X values are all zero, and each X term denotes values of included predictors. For example, X_{1i} could be the age of the subject in years, X_{2i} could be 1 for males and 0 for females, and so on. Each β denotes the change in mean outcome value (i.e. the mean difference) for each 1-unit increase in the corresponding predictor, after adjusting for other predictors. The error term, e_i, represents the residuals, and these are assumed to follow a normal distribution with a mean of zero and variance of σ^2.

After fitting this regression model using traditional methods (e.g. ordinary least squares), to adjust for overfitting a global (uniform) shrinkage factor (S) can be applied to all estimated predictor effects ($\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$, etc). That is, when making predictions in new individuals we can use the modified equation of,

$$E(Y_i) = \alpha^* + S(\hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots) \qquad (2)$$

where α^* is the revised intercept, which is re-estimated to ensure the overall predicted mean agrees with the observed mean in the development dataset (for details on how to do this, see **Table 1** and Harrell[3]). Compared to the original (non-penalised) model, this will shrink predicted values in new individuals away from the extremes, and move them towards the mean.

Implementation of this global shrinkage approach requires S to be estimated. A popular approach post-estimation is bootstrapping.[13] An alternative is to utilise the closed-form solution of Copas,[10][14]

$$S_C = 1 - \frac{(p-2)}{\text{LR}}$$

$$(3)$$

which for linear regression (see equation (8.5) in the Copas paper[10]) provides an unbiased estimate of the shrinkage factor. Here, p is the total number of predictor parameters (assumed ≥ 2) and LR is the likelihood ratio (chi-squared) statistic for the model, which can be defined as

$$\text{LR} = -2(\ln L_{null} - \ln L_{model}), \qquad (4)$$

where $\ln L_{null}$ is the log-likelihood of a model with no predictors (i.e. intercept-only null model), and $\ln L_{model}$ is the log-likelihood of the final developed model. In Section 2.2 we also show how LR can be expressed in terms of R^2 (see equation (5)).

This shrinkage estimate of S_C relates to a model developed without any variable selection procedure, and thus p represents the entire set of predictor parameters in the model. When

selection procedures are used during model development, p will be closer to the number of parameters based on the entire set of candidate predictors (i.e. all those considered for inclusion, regardless of whether they were all included in the final model).[3] Therefore, we generally define p as the total number of predictor parameters *considered* within the model development. Note that if a predictor is categorical with three of more categories, or continuous and modelled as a non-linear trend, then it will contribute two or more parameters. For these reasons, we refer to subjects per predictor parameter, rather than subjects per variable, in this article.

Example of a global shrinkage factor

For illustration, we used data from a randomised trial of 262 hypertension patients to develop a linear regression model for predicting systolic blood pressure (SBP) at the end of treatment.[15] We forced inclusion of seven predictors: age, sex, treatment group (treatment/control), smoker (yes/no), BMI, baseline SBP, and baseline diastolic blood pressure (DBP). These correspond to seven predictor parameters (i.e. p = 7). The model parameter estimates are shown in **Table 1**, and the LR statistic was 69.295. The corresponding global shrinkage factor estimate from equation (3) is:

$$S_C = 1 - \frac{(p-2)}{\text{LR}} = 1 - \frac{(7-2)}{69.295} = 0.928$$

We also used 5000 bootstrap samples to estimate the global shrinkage factor (as described elsewhere[5 16]), and this gave a very similar value of 0.94. Also the adjusted R^2 was 0.21, which is about 0.91 times the apparent R^2 value of 0.23. Therefore, even though there was no automated predictor selection based on p-values, there is still some evidence of overfitting (and thus optimism in apparent predictive performance). For a more robust prediction of SBP in new individuals, **Table 1** also shows the original beta coefficients multiplied by 0.928, which shrink the model's predictions toward the overall mean.

Table 1 Example of global shrinkage applied to a traditional (non-penalised) linear regression model for predicting systolic blood pressure at the end of treatment in hypertension patients

	Developed model	Final model after adjustment for overfitting$^\$$
Intercept	$\widehat{\alpha}$	α^*
	28.096	39.057
Predictor	$\widehat{\beta}$	$S\widehat{\beta}$ = 0.928$\widehat{\beta}$
SBP at baseline	0.462	0.429
DBP at baseline	0.411	0.381
BMI	0.013	0.012
Age	0.450	0.418
Sex	-2.050	-1.902
Treatment	-17.807	-16.525
Smoker	-2.082	-1.932

$^\$$ The revised intercept is obtained by $\alpha^* = (1 - S_C)\overline{Y} + S_C\widehat{\alpha}$ where $\overline{Y}$ is the mean outcome value in the development dataset and $\widehat{\alpha}$ is the estimated intercept from a traditional (non-penalised) model. For further details see Harrell.[3]

2.2 Shrinkage expressed in terms of sample size and R^2

We now propose utilising the Copas shrinkage factor, S_C, to inform sample size calculations at the start of a study; i.e. before individual participant data have been obtained. Specifically, we derive an expression that allows the researcher to identify the sample size and number of predictor parameters that gives an expected value of S_C close to 1 (e.g. 0.9). Our approach specifically builds on Harrell et al.,[3][6] who shows how – *after* the development dataset is obtained and a model fitted including all predictors– the shrinkage estimate can inform whether to reduce the number of predictors (using so-called data reduction techniques). Our premise is the same, except we focus on calculating the *expected* shrinkage *before* data collection, to inform sample size calculations for a new study.

We start by re-expressing S_C in terms of sample size (n), number of predictor parameters (p), and R^2, the proportion of variability explained. Let R^2_{app} be the apparent estimate of a prediction model's R^2 in the dataset used to develop the model. That is, $R^2_{app} = 1 - (\widehat{\sigma}^2_{model}/\widehat{\sigma}^2_{null})$, and thus $0 \leq R^2_{app} \leq 1$. As shown elsewhere,[17][18] the LR statistic can be expressed in terms of the sample size and R^2_{app}, as follows:

$$\text{LR} = -n \ln(1 - R^2_{app}) \tag{5}$$

Applying equation (5) within equation (3), the Copas shrinkage formula becomes,

$$S_C = 1 + \frac{p - 2}{n \ln(1 - R^2_{app})}$$

(6)

Equation (6) cannot be used to directly inform the sample size (n) in advance of model fitting, because R^2_{app} is a *post*-estimation measure of model fit. However, an approximately unbiased (optimism-adjusted) estimate of the proportion of variation explained is R^2_{adj},[10][19]

$$R^2_{adj} = 1 - \left((1 - R^2_{app})\frac{(n - 1)}{(n - p - 1)}\right) = \frac{(n - 1)R^2_{app} - p}{(n - p - 1)}$$

(7)

and rearranging gives,

$$R^2_{app} = \frac{R^2_{adj}(n - p - 1) + p}{(n - 1)}$$

(8)

Therefore, applying equation (8) within equation (6) provides:

$$S_C = 1 + \frac{p - 2}{n \ln\left(1 - \left(\frac{R^2_{adj}(n - p - 1) + p}{(n - 1)}\right)\right)}$$

(9)

Hence, we now have an expression for the expected shrinkage factor conditional on a particular R^2_{adj} and, crucially, the sample size, n, and number of predictor parameters, p. When studying equation (9), we observe that the expected shrinkage will decrease (i.e. S_C will move closer to 1) as n increases, as p decreases, as n/p increases, and as R^2_{adj} increases. Therefore, shrinkage (overfitting) will be a larger concern is development datasets with a small number of subjects, a large number of predictor parameters (relative to the number of subjects), and when the proportion of variance explained by the model is low.

2.3 Criterion (i): calculating sample size to ensure a shrinkage factor ≥ 0.9

Recall, at the start of Section 2 we explained that it is important to minimise the potential for overfitting. Therefore, when designing a new model development study, we propose researchers should utilise equation (9) to reveal the sample size (n) needed to obtain a targeted value of S_C. We suggest using a value of $S_C \geq 0.9$, such that predictor effects would shrink by $\leq 10\%$, which represents small overfitting. This is in accordance with Harrell,[3] who suggests that if the shrinkage estimate "falls below 0.9, for example, we may be concerned with the lack of calibration the model may experience on new data".

Although there is no closed form solution for n based on equation (9), an iterative process can be used to identify the value of n that gives the desired S_C conditional on a chosen p and R_{adj}^2. For example, to obtain an expected S_C of 0.9 for a hypothetical model with up to 30 predictor parameters and an anticipated R_{adj}^2 of 0.7, a sample size of 206 subjects is required to meet criterion (i) as:

$$S_C = 1 + \frac{30 - 2}{206 \ln\left(1 - \left(\frac{0.70(\ 206 - 30 - 1) + 30}{(206 - 1)}\right)\right)} = 0.90$$

This equates to 206/30 = 6.87 subjects per predictor parameter; that is, about 6.87 subjects are required for each predictor parameter considered. If it was rather considered that up to 50 predictor parameters are needed for an anticipated R_{adj}^2 of 0.7, then a sample size of 355 subjects is required to obtain an expected S_C of 0.9, corresponding to 7.10 subjects per predictor parameter. Hence, the number of subjects per predictor parameter changes depending on the number of predictor parameters considered.

In situations where the calculated sample size is considered unrealistic (e.g. due to time and cost constraints), p could be lowered by reducing the number of candidate predictor parameters. For example, those predictors known from previous studies (or systematic reviews) to have predictive value could be prioritised, or two or more predictors could be combined into one, such as BMI instead of weight and height. Alternatively after data collection, unsupervised learning techniques such as principal component analysis could be used, which are blind to the outcome values. In the above example, reducing the number of predictor parameters to 25 leads to a sample size of 169 subjects to meet criterion (i):

$$S_C = 1 + \frac{25 - 2}{169 \ln\left(1 - \left(\frac{0.70(\,169 - 25 - 1) + 25}{(169 - 1)}\right)\right)} = 0.90$$

Thus, by removing five predictor parameters the required sample size to meet criterion (i) is reduced by 37 subjects.

When using a more stringent shrinkage factor, say of 0.95, then the necessary sample size to meet criterion (i) will be increased substantially. For instance, in the previous example with 25 predictor parameters and an anticipated R^2_{adj} of 0.7, the sample size required is increased from 169 to 361 when increasing S_C from 0.90 to 0.95. Hence, over twice the sample size is needed to reduce the expected shrinkage from 10% to 5%. For this reason, we anticipate that an S_C of 0.90 will often be a pragmatic choice for criterion (i).[3]

2.4   Criterion (ii): calculating sample size to ensure a small absolute difference in R^2_{adj} and R^2_{app}

Criterion (i) focuses on shrinkage of predictor effects, which is a multiplicative measure of overfitting (i.e. on the relative scale), and therefore Harrell also suggests to evaluate overfitting on the absolute scale.[3] To address this, our second criterion for minimum sample size is to ensure the difference (δ) between R^2_{app} and R^2_{adj} is small, say ≤ 0.05, such that the optimism in the developed model's apparent proportion of variance explained is small.

Utilising equation (8), the difference in R^2_{app} and R^2_{adj} can be written as:

$$\delta = R^2_{app} - R^2_{adj}$$
$$= \frac{R^2_{adj}(\,n - p - 1) + p}{(n - 1)} - R^2_{adj}$$
$$= \frac{R^2_{adj}(\,n - p - 1) + p - (n - 1)R^2_{adj}}{(n - 1)}$$
$$= \frac{-pR^2_{adj} + p}{(n - 1)}$$
$$= \frac{p(1 - R^2_{adj})}{(n - 1)}$$

After rearranging this solution, we find that to meet criterion (ii) we require the number of subjects to be,

$$n \geq 1 + \frac{p(1 - R_{adj}^2)}{\delta}$$

<div align="right">(10)</div>

where δ is a small value, such as ≤ 0.05. For example, returning to our hypothetical model with an anticipated R_{adj}^2 of 0.7 and up to 30 potential predictor parameters, the sample size required to meet criterion (ii) is:

$$n \geq 1 + \frac{p(1 - R_{adj}^2)}{\delta} = 1 + \frac{30(1 - 0.7)}{0.05} = 181$$

This is slightly lower than the sample size of 206 identified for criterion (i) in Section 2.3.

Equation (10) reveals that, similar to the shrinkage approach for criterion (i), the required sample size for criterion (ii) will increase as p increases and R_{adj}^2 decreases. For example, if our hypothetical model had an anticipated R_{adj}^2 of 0.3 rather than 0.7, then the sample size to meet criterion (ii) increases substantially to 421.

2.5 How to pre-specify R_{adj}^2

To identify a sample size to meet our criteria (i) and (ii), researchers have to pre-specify a value for the model's anticipated R_{adj}^2. How should this be done? We recommend identifying previous prediction model studies for the same or similar populations and outcomes of interest, and extracting their R_{adj}^2 values, which are usually well-reported for linear regression models. Helpful for this purpose are systematic reviews of existing models,[20] and registries that record the prediction models available in a particular field.[21] If only an R_{app}^2 value is reported in a model development study, then its R_{adj}^2 can be derived using equation (7) as long as the study's n and p can also be obtained. Note that if R_{app}^2 is reported from an external validation study of an existing model, there is no need for adjustment (i.e. $R_{app}^2 = R_{adj}^2$), as the validation dataset provides a direct estimate of the model's performance in new individuals (free from overfitting concerns as there is no model development therein). In other words, deriving R_{adj}^2 based on a reported R_{app}^2 is necessary when the latter is from a model development study, but not when it is from an appropriate external validation study of an existing model.[5] Guidance for choosing an R_{adj}^2 value in the absence of any prior information is given in the Discussion.

3   Sample size required for precise estimation of the residual standard deviation and mean predicted outcome value

In addition to reducing the potential for overfitting, Harrell notes that sample sizes should be large enough to precisely estimate key model parameters such as the intercept or residual variance.[3] We now address this.

3.1   Criterion (iii): precise estimate of the residual standard deviation

A precise estimate of the residual standard deviation is essential, as it is subsequently used to estimate R^2 , and also to derive the standard errors and confidence intervals for the intercept and predictor effects (betas). For simplicity, Harrell suggests focusing on the standard deviation, σ_{null} say, in a null model (i.e. intercept-only model),[3] and ensuring the lower and upper bounds of a 95% confidence interval for σ_{null} have a small multiplicative margin of error (MMOE) around the true value of σ_{null}. Assuming residuals are approximately normally distributed, this approach can be extended to consider the MMOE for estimating σ_{model}, the residual standard deviation in the developed prediction model, by:

$$\text{MMOE} = \sqrt{\max\left(\frac{\chi^2_{1-\frac{\alpha}{2},n-p-1}}{n-p-1}, \frac{n-p-1}{\chi^2_{\frac{\alpha}{2},n-p-1}}\right)} \tag{11}$$

where $\chi^2_{1-\frac{\alpha}{2},n-p-1}$ and $\chi^2_{\frac{\alpha}{2},n-p-1}$ are the critical values of a χ^2 distribution with $n-p-1$ degrees of freedom for which there is, respectively, a probability of $1-\frac{\alpha}{2}$ and $\frac{\alpha}{2}$ of being less than the critical value. The second term within the bracket of equation (11) will typically give the largest MMOE.

For example, consider that we wish to ensure (with 95% confidence) that the margin of error is within 20% of the true value, i.e. $1.0 \leq \text{MMOE} \leq 1.2$. For a null model (i.e. one containing no predictors), then equation (11) reveals that a sample size of at least 70 subjects is needed to meet this criterion,[3] as this gives an MMOE of 1.2. Therefore, in a multivariable model with p predictor parameters, the minimum sample required to meet a MMOE ≤ 1.2 for criterion (iii) is simply $70 + p$.

However, we recommend a more stringent margin of error of within 10% of the true value, i.e. $1.0 \leq \text{MMOE} \leq 1.1$. In a null model, equation (11) reveals that a sample size of at least 234 subjects is needed to ensure a MMOE ≤ 1.1. Therefore, in a multivariable model with p predictor parameters, the minimum sample required to meet a MMOE of ≤ 1.1 for criterion (iii) is simply $234 + p$.

To illustrate this, let us return to the hypothetical example initially described in Section 2.3, where the sample size to meet criterion (i) was 206 and for criterion (ii) was 181. However, these values correspond to a MMOE of greater than 1.1 for σ_{model}. For example, taking the sample size of 206 subjects and 30 predictor parameters, this corresponds to a $\chi^2_{0.975,206-30-1}$ = 213.52 and $\chi^2_{0.025,206-30-1}$ = 140.26, and thus (using equation (11)) a MMOE of $\sqrt{(206-30-1)/140.26} = 1.12$, and 12% potential margin of error in the estimate of σ_{model}.

Rather, with 30 predictor parameters, to achieve an expected MMOE of 1.1 we require $234 + p = 234 + 30 = 264$ subjects, because the maximum value of equation (11) is then exactly 1.10:

$$\sqrt{(264-30-1)/\chi^2_{0.025,264-30-1}} = \sqrt{(264-30-1)/192.615} = 1.10$$

Hence, in this example the minimum sample size of 264 subjects for criterion (iii) (i.e. to ensure a MMOE ≤1.1) is more stringent than those identified for criteria (i) and (ii).

3.2 Criterion (iv): precise estimate of the mean predicted outcome value (model intercept)

It is also important for model predictions to be precise; in particular, it is fundamental that the mean predicted outcome value is precisely estimated. If we assume our model will include predictors centred at their mean values in the developed dataset, then the fitted model's intercept ($\hat{\alpha}_{model}$) will correspond to the predicted outcome value for an individual with mean predictor values. This estimate will be similar (though not identical) to the overall mean outcome in the population of interest; such a population mean estimate has variance of $\hat{\sigma}^2_{null}/n$. However, in a linear regression model with multiple predictors the residual variance is $\hat{\sigma}^2_{model}$, and so the fitted model's intercept will have an approximate variance of:[Footnote1]

$$var(\hat{\alpha}_{model}) = \hat{\sigma}^2_{model}/n \approx \sigma^2_{null}(1 - R^2_{adj})/n$$

(Footnote1: We use R^2_{adj} in this equation rather than R^2_{app} to be conservative)

Then, a 95% confidence interval for the model intercept is,

$$\hat{\alpha}_{model} \pm \left(t_{1-\frac{0.05}{2},n-p-1}\sqrt{\frac{\sigma^2_{null}(1-R^2_{adj})}{n}} \right)$$

(12)

where $t_{1-\frac{0.05}{2},n-p-1}$ is the critical value of a t-distribution with $n - p - 1$ degrees of freedom for which there is a probability of $1 - \frac{0.05}{2}$ below the critical value. Therefore, to derive this confidence interval in advance of model development, the researcher needs to pre-specify (e.g. from previous studies) sensible values for the anticipated mean outcome value ($\hat{\alpha}_{model}$), the population (null model) variance (σ^2_{null}), and R^2_{adj}. Then, the researcher can identify the sample size that ensures a sufficiently narrow confidence interval for α_{model}, to satisfy criterion (iv). For example, they might ensure the lower and upper bounds are within a small multiplicative margin of error (MMOE) of the anticipated prediction mean (i.e. $1.0 \leq$ MMOE ≤ 1.1).

However, what constitutes a sufficiently narrow confidence interval will be context specific. A sensible start point is to examine the confidence interval width when using the sample sizes identified for criterion (i) to (iii). For example, let us return to our hypothetical model with an anticipated R^2_{adj} of 0.7 and p = 30 predictor parameters, and now assume that the target population has an anticipated mean blood pressure ($\hat{\alpha}_{model}$) of 165 and variance ($\hat{\sigma}^2_{null}$) of 18^2. Then, using equation (12) and a sample size of 264 subjects identified by criterion (iii), the 95% confidence interval for the mean predicted outcome value is:

$$165 \pm \left(t_{0.975,(264-30-1)} \sqrt{\frac{18^2(1 - 0.7)}{264}} \right) = 163.8 \text{ to } 166.2$$

This is reassuringly precise, with the upper bound just 1.2 higher than the true mean of 165; this corresponds to a margin of error within 10% of the true mean (indeed, MMOE is 166.2/165 = 1.007, and thus margin of error < 1%).

4 Worked example: Prediction of lung function in African-Americans

A step-by-step summary of our sample size proposal is given Figure 1, and we now apply it to a worked example. Kumar et al. use linear regression to identify predictors of lung function (i.e. the forced expiratory volume in 1 second, FEV1) in African-American participants.[22] Let us assume that we want to build on this work, by formally developing a linear regression model to predict FEV1 in African-American women. The aim could be to flag those individuals with low FEV1 values, as these are at risk of chronic obstructive pulmonary disease (COPD). We now go through the sample size calculation process.

Figure 1: Summary of the steps involved in our sample size calculation for developing a multivariable prediction model

- **STEP 1**: **Choose the number of candidate predictors of interest for inclusion in the model, and calculate the corresponding number of predictor parameters (p)**. Recognise that one predictor may require two or more parameters; for example, a k category predictor requires *k-1* parameters, and a continuous predictor modelled with a non-linear trend requires > 1 parameter to be estimated. Also include any potential interaction terms toward the total p.

- **STEP 2**: **Choose a value for the anticipated proportion of variance explained (R^2) for the new model** (referred to as R^2_{adj} in our article). For example, the R^2_{adj} value for a previously published model in the same setting and population could be used as a lower bound for the anticipated R^2 of the new model.

- **STEP 3**: **Criterion (i)** – use equation (9) to calculate the minimum sample size required to ensure Copas' global shrinkage factor (S_C) is close to 1. We generally recommend a value of $S_C \geq 0.90$, which reflects a small amount of overfitting during model development.

- **STEP 4**: **Criterion (ii)** – use equation (10) to calculate the minimum sample size required to ensure a small absolute difference of ≤ 0.05 in the developed model's R^2_{adj} and R^2_{app}.

- **STEP 5**: **Criterion (iii)** – based on equation (11), calculate the minimum sample size required to ensure a precise estimate of the residual standard deviation (σ_{model}). We generally recommend at least $234 + p$ subjects, which ensures the σ_{model} estimate has no more than a 10% margin of error from the true value.

- **STEP 6**: **Criterion (iv)** – based on equation (12), calculate the minimum sample size required to ensure a precise estimate of the mean predicted outcome value in the developed model (precise model intercept in a model with mean-centred predictors). This requires the researcher to pre-specify (e.g. from previous studies) sensible values for the anticipated mean outcome value ($\hat{\alpha}_{model}$) and the population (null model) variance (σ^2_{null}), in addition to R^2. What constitutes a precise estimate is context specific, but a broad suggestion is to at least ensure a confidence interval with lower and upper values within a 10% multiplicative margin of error from the true mean.

- **STEP 7: Final sample size** - the required minimum sample size is the maximum value from steps 3 to 6, to ensure that each of criteria (i) to (iv) are met. Researchers might also examine whether the sample size would give precise estimates of key predictor effects (see Section 5).

  If the calculated sample size is not considered achievable due to criteria (i), (ii) or (iv), consider reducing the number of candidate predictors (and thus *p*) to reduce the required sample size (whilst still meeting criterion (iii)). For example, prioritise those predictors identified as important from existing systematic reviews, or consider data reduction techniques such as principal component analysis (blinded to predictor-outcome associations in the development dataset). We do not recommend reducing the size of S_C or increasing R^2_{adj}.

4.1 Step-by-step application

Steps 1 and 2: Identifying a value (lower bound) for the model's anticipated R^2_{adj} and choosing p

Kumar et al. (see their supplementary material[22]) report the performance of a model for prediction of FEV1 in women, with the model containing three predictors (age, height, African ancestry) and four predictor parameters. The R^2_{adj} was on average about 0.2 across three different datasets. Therefore, we could use this value as a lower bound for the anticipated R^2_{adj} in the new model. Also, let us assume that there will be up to 25 predictor parameters in this new model (including the four used in the original model), and thus p = 25.

Step 3: Criterion (i) – ensuring S_C is close to 1

Based on the chosen R^2_{adj} = 0.2 and p = 25, to ensure an expected S_C of 0.9, a sample size of 918 subjects is needed because (using equation (9)):

$$S_C = 1 + \frac{p - 2}{n \ln\left(1 - \left(\frac{R^2_{adj}(n - p - 1) + p}{(n - 1)}\right)\right)}$$

$$= 1 + \frac{25 - 2}{918 \ln\left(1 - \left(\frac{0.2(918 - 25 - 1) + 25}{(918 - 1)}\right)\right)}$$

$$= 0.90$$

This corresponds to requiring 36.7 subjects per predictor parameter to meet criterion (i).

Figure 2 shows how the expected shrinkage, S_C, derived from equation (9) changes according to n and p, conditional on an R^2_{adj} of 0.2. As n increases and p decreases, the expected S_C becomes closer to 1. Furthermore, for an S_C above 0.9, very large increases in the sample size are needed to improve the expected S_C. For example, if we wanted to use a more stringent criteria for low overfitting of S_C = 0.95 then a sample size of 1949 subjects is required (78 subjects per predictor parameter), which is over double the number when S_C is 0.9.

Step 4: Criterion (ii) – ensuring small absolute difference between R^2_{adj} and R^2_{app}

Using equation (10), we can calculate the minimum sample size needed to ensure the absolute difference between R^2_{adj} and R^2_{app} is 0.05 or less:

$$n \geq 1 + \frac{p(1 - R^2_{adj})}{\delta} = 1 + \frac{25(1 - 0.2)}{0.05} = 401$$

Therefore, at least 401 subjects are required to ensure a small absolute magnitude of overfitting based on the difference between R^2_{adj} and R^2_{app}.

Step 5: Criterion (iii) – ensuring a precise estimate of the residual standard deviation
When p = 25, then equation (11) reveals that we need at least 259 subjects (= $234 + p$) to ensure a margin of error of $\leq 10\%$ in the estimate of the model's residual standard deviation, σ_{model}. This is because the maximum value of equation (11) is:

$$\sqrt{(259 - 25 - 1)/\chi^2_{0.025,259-25-1}} = \sqrt{(259 - 25 - 1)/192.615} = 1.10$$

Step 6: Criterion (iv) – ensure a precise estimate of the mean predicted outcome (model intercept)
Based on the results in Kumar et al.[22] the population mean and variance of FEV1 are about 1.90 litres and 0.6^2, respectively. Based on these values and using equation (12) with an assumed sample size of 918 subjects (as needed for criterion (i)), we obtain a 95% confidence interval for the predicted mean outcome value (model intercept when predictors are mean-centered) of:

$$1.90 \pm \left(t_{0.975,(918-25-1)} \sqrt{\frac{0.6^2(1 - 0.2)}{918}} \right) = 1.87 \text{ to } 1.93$$

This is reassuringly precise, and the upper bound indicates a multiplicative margin of error <1.1 (i.e. within 10%) of the true mean outcome value.

Step 7: Identify sample size that ensures all criteria are met
Based on the largest sample size calculations identified in steps 3 to 6, the final minimum sample size required is 918 subjects. This is driven by criterion (i), to ensure an expected shrinkage factor $\geq$ 0.9.

4.2   What if the sample size is not considered achievable?

If the sample size of 918 subjects was not considered achievable (e.g. due to time or cost constraints), then what should be done? For criterion (i), we do not recommend reducing S_C below 0.9, as our main premise is to minimise overfitting. That leaves two other potential options: use a larger R^2_{adj} value or reduce p. We do not recommend the first option. The R^2_{adj} values reported in previous articles are themselves only estimates, and it is hard to judge how much R^2_{adj} will be improved in a new model. It is far better to be conservative in the choice of R^2_{adj} and adopt larger

sample sizes, rather than naively aiming for smaller sample sizes that ultimately may not be fit for purpose.

Therefore, the best approach is to reduce the number of candidate predictor parameters, p. Returning to the hypothetical example where the assumed R^2_{adj} is 0.2, a reduced set of 15 predictor parameters would lower the sample size required for criterion (i) to 515 subjects (Figure 2). Thus, after sacrificing 10 parameters by removing some predictors, the researchers requires 403 fewer subjects to target an S_C of 0.9. The choice of which predictors to prioritise could be based on external evidence (e.g. from systematic reviews) and, after data collection, data reduction techniques such as principal components analysis (which are based on observed correlation amongst predictors only, and not observed predictor-outcome associations). All those predictors within the existing Kumar et al. model are best retained in the model, to justify the assumption that R^2_{adj} is at least 0.2. A sample size of 515 subjects still ensures criteria (ii) to (iv) are met. For example, for criterion (iv), the width of the 95% confidence interval for the prediction mean would still be very narrow (1.85 to 1.95).

Figure 2: Expected shrinkage (S_C) from equation (9) for a prediction model of lung function in African-Americans, conditional on a particular sample size (n), number of predictor parameters (p), and an assumed R^2_{adj} of 0.2
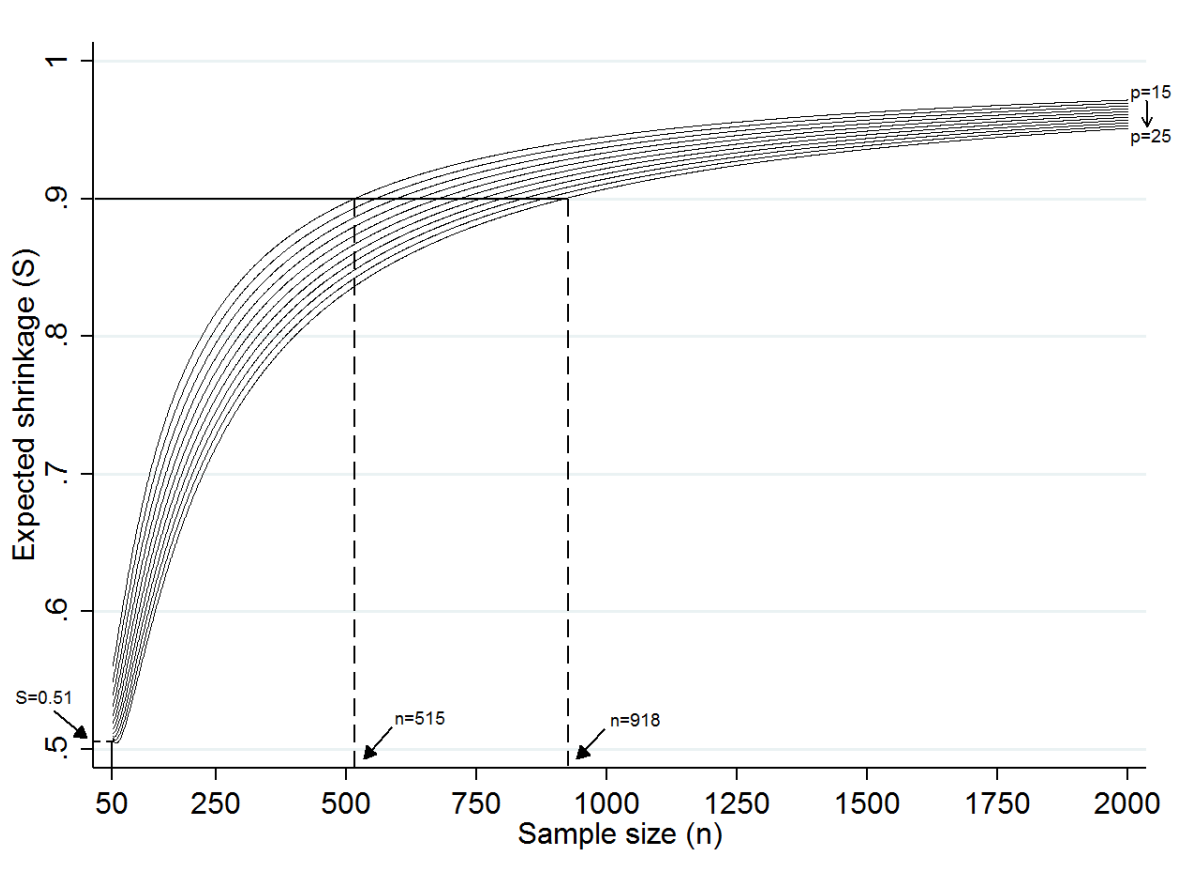
4.3 Comparison to other suggested sample size proposals

We now contrast our derived sample size of 918 subjects to those from two other suggested sample size approaches for linear regression models. Though not intended for informing prediction model development, a recent recommendation suggests two subjects per predictor parameter for adequate estimation of predictor effects in linear regression.[23] In our example, two subjects for each of the 25 predictor parameters leads to a substantially smaller sample size of 50 subjects. However, using equation (9), this corresponds to an expected shrinkage of S_C = 0.51, which reflects substantial overfitting and does not meet criterion (i) (Figure 2).

Alternatively, Harrell suggests there are at least 15 subjects per predictor parameter (see Chapter 4 in his book[3]), which in this example implies a sample size of at least 375. However, using equation (9), a sample size of 375 subjects corresponds to an expected shrinkage of S_C = 0.79, which still suggests large overfitting.

5 Potential additional criteria

Criteria (i) to (iv) form our main proposal for the minimum sample size required when developing a prediction model for continuous outcomes. However, we now briefly mention two additional criteria that may also be important to consider.

5.1 Ensuring precise estimation of R^2_{adj} and the mean-square error

Criterion (ii) ensures that there is a small absolute difference between R^2_{adj} and R^2_{app}, to reflect low overfitting. A related concept is to ensure a precise confidence interval for R^2_{adj}.[24] Tan gives an excellent overview of various exact and approximate approaches to calculate a confidence interval for R^2_{adj},[25] given a developed model's R^2_{app}, n and p. For example, Lee proposes a confidence interval based on a scaled non-central F distribution approximation to the distribution of R^2.[26] This can be implemented in SAS,[27] or in R using the ci.R2 function of the MBESS package by Kelley.[28-30] Furthermore, the ss.aipe.R2 function within MBESS identifies the sample size required to ensure Lee's confidence interval is sufficiently narrow.

We applied the ss.aipe.R2 function to the lung function model described in Section 4. This identified that 835 subjects are required to ensure the expected width of the confidence interval for R^2_{adj} is exactly 0.10, assuming R^2_{adj} is 0.20 and p = 25. This sample size is lower than the 918 subjects required to meet criterion (i), and hence 918 subjects is still the minimum sample size required.[Footnote2]

Footnote2: With 918 subjects, and assuming R^2_{adj} is 0.20 and p = 25, Lee's 95% confidence interval for R^2_{adj} is 0.154 to 0.250.

Ensuring precise estimates of R^2_{adj} and the residual standard deviation (criteria (iii)) also helps ensure a precise estimate of the mean-square error (MSE) of the model's predicted outcome values, as $\text{MSE} = \sigma^2_{null}(1 - R^2_{adj})$.

5.2   Ensuring precise estimation of key predictor effects

Criterion (iv) ensures a precise estimate of the mean predicted outcome value in the entire target population. Ideally, predictions should also be precise across the entire spectrum of predicted values, not just at the mean. This is challenging, but is helped by ensuring the effects of key predictors are estimated precisely. The precision of a particular predictor effect in a fitted linear regression model depends on the sample size, the estimated residual variance, the correlation of the predictor with other included predictors, and the variance of the predictor values.[31] For brevity, we do not consider this in detail here, and refer the reader to other articles that focus on this.[31-35] In particular, the ss.aipe.rc function with the MBESS package identifies the sample size required to ensure the confidence interval around a predictor's effect is sufficiently narrow.[30 33 35]

Returning to the lung function example, let us consider that our new model will potentially include smoking as a predictor, defined as a binary variable (current/previous smokers versus non-smokers). Furthermore, assume that the mean difference in FEV1 for smokers and non-smokers is -0.5, and that (based on Kumar et al.[22]) 50% of subjects will be current/previous smokers. Also, assume (conservatively) that the final model will have R^2_{adj} of 0.2 and that the correlation is 0.5 between smoking and other included predictors. Using the the ss.aipe.rc function in R, we identify that 619 subjects are required to ensure a confidence interval width of 0.2 (and thus the lower and upper bounds are within 0.1 of the true value of -0.5). This is reassuring, and again less than the 918 subjects required to satisfy criterion (i).

Precise estimation of predictor effects will be especially difficult for those predictors with smallest variance in their values, as their confidence intervals are likely to be widest.[31] In particular, categorical predictors with low prevalence in certain categories have small variances.[36] In our example had the percentage of smokers been assumed to be, say, 10% rather than 50%, then repeating the calculation identifies that 1668 subjects are needed for a confidence interval width of 0.2. In this situation, we would need to increase the sample size beyond 918 subjects previously identified to meet criteria (i) to (iv), or justify relaxing the magnitude of precision desired. For

example if we were willing to widen the expected confidence interval width to 0.3, then this considerably reduces the number of required subjects to 757, but the interval is still fairly precise and all well below zero (-0.65 to -0.35).

6 Discussion

Sample size calculations are a fundamental part of designing a study to develop a new prediction model. In this article, we proposed four criteria to identify the minimum sample size needed to minimise overfitting whilst ensuring precise estimates of key model parameters. Criterion (i) forms the most novel aspect of our sample size proposal, as it allows researchers to identify n and p that correspond to an expected shrinkage factor close to 1, such as 0.9, which reflects low overfitting. Furthermore, it allows the sample size to be tailored to each model of interest, through the pre-specification of the anticipated percentage of variation explained, R^2_{adj}, which is a measure of overall model fit. The chosen value of R^2_{adj} strongly influences the amount of shrinkage required, with larger values requiring less shrinkage (with other things, such as p, being equal).[37] This issue is currently ignored when using blanket rules of thumb for sample size.

Researchers should use previous evidence from other prediction model's in the same setting to ascertain a (conservative) value for the new model's potential R^2_{adj} value. If no relevant prediction models exist, then information from predictor finding studies (e.g. studies aiming to estimate the prognostic effect of a particular predictor adjusted for other existing factors[4]) might be relevant. Even though such studies are primarily focused on the estimation of the effect of a particular predictor, they typically involve multivariable modelling and therefore often also report R^2_{app} and R^2_{adj} values. Where truly no prior information exists about the potential R^2_{adj} value, researchers should recognise that medical diagnosis and prediction of health-related outcomes are, generally speaking, low signal:noise ratio situations. It is not uncommon in these situations to see R^2_{adj} values in the 0.1 to 0.2 range. Therefore, in the absence of other information, we suggest that sample sizes be derived assuming $R^2_{adj} = 0.15$. An exception is when predictors include 'direct' (mechanistic) measurements, such as the baseline version of the continuous outcome (e.g. when predicting lung function one year after measuring baseline lung function). Then, in this special situation, an $R^2_{adj} = 0.5$ may be a more appropriate default choice.

In practice, *after* a model development dataset is obtained, a better approach for estimating the shrinkage factor is to use a resampling approach such as bootstrapping.[10][16] However, as our sample

size calculations are focused on situations before any data collection, it is not possible to incorporate such a resampling approach. In situations where a development dataset is already available, containing a specific number of subjects and predictors, our approach could be used to identify whether a reduction in the number of predictors is needed (prior to beginning the modelling). Indeed, Harrell previously illustrated this concept by using the shrinkage estimate from the full model (including all predictors) to gauge whether the number of predictors should be reduced.[3] This could then incorporate bootstrapping (rather than the Copas formula) to estimate the shrinkage. However, this should be done blind to the estimated predictor effects, as otherwise the decisions about inclusion are already being made based on the full set of predictors. Similarly, when planning to use a predictor selection method (such as backwards selection) during model development, researchers should define p as the total number of parameters due to all predictors considered (screened), and not just the subset that are included in the final model.[5] As Harrell notes,[3] the value of p should be honest.

A potential limitation of our work is that multiple sample size calculations are required, to address each of the criteria considered. However, this reflects the different elements that require consideration when developing a prediction model. Criteria (iii) and (iv) are needed to ensure that there will be a small margin of error in the estimates of the residual standard deviation and the mean predicted outcome value (model intercept). This is often overlooked when considering the sample size. In particular, at least $234 + p$ subjects are always required to ensure a multiplicative margin of error (MMOE) of ≤ 1.1 for estimating the model's residual standard deviation, σ_{model}. Section 5 emphasised that further criteria may also be needed in some settings. In particular, ensuring precise estimates of predictor effects may be important, especially in settings where key predictors have low variance (e.g. categorical predictors with few subjects in certain categories).

In summary, we have proposed how to ascertain the minimum sample size needed to develop a prediction model using linear regression. We hope this encourages researchers to move away from rules of thumb, and to rather focus on attaining sample sizes that ensure precise estimates and reduce the potential for overfitting, in order to develop more robust prediction models. We are currently writing software modules to implement the approach. Our accompanying paper extends the work to binary and time-to-event outcomes.[7]

Reference List

1. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10(2):e1001381.

2. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer 2009.

3. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis (Second Edition). New York: Springer 2015.

4. Riley RD, Hayden JA, Steyerberg EW, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med* 2013;10(2):e1001380.

5. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162(1):W1-73.

6. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15(4):361-87.

7. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* (in-press).

8. Van Houwelingen JC. Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Statistica Neerlandica* 2001;55:17-34.

9. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. *Stat Med* 1990;9(11):1303-25.

10. Copas JB. Regression, Prediction and Shrinkage. *J Royal Statist Soc B* 1983;45(3):311-54.

11. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Statist Soc B* 1996;58:267-88.

12. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. *BMJ* 2015;351:h3868.

13. Efron B. Bootstrap methods: Another look at the jackknife *The Annals of Statistics* 1979;7(1):1-26.

14. Copas JB. Using regression models for prediction: shrinkage and regression to the mean. *Stat Methods Med Res* 1997;6(2):167-83.

15. Riley RD, Kauser I, Bland M, et al. Meta-analysis of randomised trials with a continuous outcome according to baseline imbalance and availability of individual participant data. *Stat Med* 2013;32(16):2747-66.

16. Steyerberg EW, Harrell FEJ, Borsboom GJ, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001;54:774-81.

17. Magee L. R-squared Measures Based on Wald and Likelihood Ratio Joint Significance Tests. *The American Statistician* 1990;44(3):250-53.

18. Hendry DF, Nielsen B. Econometric Modeling: A Likelihood Approach: Princeton University Press 2012.

19. Goldberger AS. Econometric Theory. New York: Wiley 1964.

20. Debray TP, Damen JA, Snell KI, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017;356:i6460.

21. Wessler BS, Paulus J, Lundquist CM, et al. Tufts PACE Clinical Predictive Model Registry: update 1990 through 2015. *Diagnostic and Prognostic Research* 2017;1(1):20.

22. Kumar R, Seibold MA, Aldrich MC, et al. Genetic Ancestry in Lung-Function Predictions. *New England Journal of Medicine* 2010;363(4):321-30.

23. Austin PC, Steyerberg EW. The number of subjects per variable required in linear regression analyses. *J Clin Epidemiol* 2015;68(6):627-36.

24. Algina J, Olejnik S. Determining Sample Size for Accurate Estimation of the Squared Multiple Correlation Coefficient. *Multivariate Behavioral Research* 2000;35(1):119-37.

25. Tan L. Confidence Intervals for Comparison of the Squared Multiple Correlation Coefficients of Non-nested Models. *Electronic Thesis and Dissertation Repository (Paper 384)* 2012.

26. Lee YS. Tables of the Upper Percentage Points of the Multiple Correlation. *Biometrika* 1971;59:175-89.

27. Zou GY. Toward using confidence intervals to compare correlations. *Psychological methods* 2007;12(4):399-413.

28. Kelley K. Confidence Intervals for Standardized Effect Sizes: Theory, Application, and Implementation. *Journal of Statistical Software* 2007;20(8):24.

29. Kelley K. Methods for the Behavioral, Educational, and Social Sciences: An R package. *Behavior Research Methods* 2007;39(4):979-84.

30. Kelley K. MBESS (Version 4.0.0 and higher) [computer software and manual]. 2017;Accessible from https://CRAN.R-project.org/package=MBESS

31. McClelland GH. Increasing statistical power without increasing sample size. *American Psychologist* 2000;55(8): 963-64.

32. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. *Stat Med* 1998;17(14):1623-34.

33. Kelley K, Maxwell SE. Sample size for multiple regression: obtaining regression coefficients that are accurate, not simply significant. *Psychol Methods* 2003;8(3):305-21.

34. Kelley K. Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. *Behavior Research Methods* 2007;39(4):755-66.

35. Maxwell SE, Kelley K, Rausch JR. Sample size planning for statistical power and accuracy in parameter estimation. *Annu Rev Psychol* 2008;59:537-63.

36. 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.

37. Knofczynski GT, Mundfrom D. Sample Sizes When Using Multiple Linear Regression for Prediction. *Educational and Psychological Measurement* 2008;68(3):431-42.