

# Minimum sample size for developing a multivariable prediction model using multinomial logistic regression

Statistical Methods in Medical Research

2023, Vol. 32(3) 555–571

© The Author(s) 2023



Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/09622802231151220

[journals.sagepub.com/home/smm](https://journals.sagepub.com/home/smm)

Alexander Pate<sup>1</sup> , Richard D Riley<sup>2</sup>, Gary S Collins<sup>3,4</sup> ,  
Maarten van Smeden<sup>5,6</sup>, Ben Van Calster<sup>7,8,9</sup> , Joie Ensor<sup>2</sup>   
and Glen P Martin<sup>1</sup> 

## Abstract

**Aims:** Multinomial logistic regression models allow one to predict the risk of a categorical outcome with  $> 2$  categories. When developing such a model, researchers should ensure the number of participants ( $n$ ) is appropriate relative to the number of events ( $E_k$ ) and the number of predictor parameters ( $p_k$ ) for each category  $k$ . We propose three criteria to determine the minimum  $n$  required in light of existing criteria developed for binary outcomes.

**Proposed criteria:** The first criterion aims to minimise the model overfitting. The second aims to minimise the difference between the observed and adjusted  $R^2$  Nagelkerke. The third criterion aims to ensure the overall risk is estimated precisely. For criterion (i), we show the sample size must be based on the anticipated Cox-snell  $R^2$  of distinct ‘one-to-one’ logistic regression models corresponding to the sub-models of the multinomial logistic regression, rather than on the overall Cox-snell  $R^2$  of the multinomial logistic regression.

**Evaluation of criteria:** We tested the performance of the proposed criteria (i) through a simulation study and found that it resulted in the desired level of overfitting. Criterion (ii) and (iii) were natural extensions from previously proposed criteria for binary outcomes and did not require evaluation through simulation.

**Summary:** We illustrated how to implement the sample size criteria through a worked example considering the development of a multinomial risk prediction model for tumour type when presented with an ovarian mass. Code is provided for the simulation and worked example. We will embed our proposed criteria within the `pmsampsize` R library and Stata modules.

## Keywords

Clinical prediction models, sample size, multinomial logistic regression, shrinkage

<sup>1</sup>Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK

<sup>2</sup>Centre for Prognosis Research, School of Medicine, Keele University, Staffordshire, UK

<sup>3</sup>Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

<sup>4</sup>NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK

<sup>5</sup>Julius Center for Health Sciences, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

<sup>6</sup>Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, Netherlands

<sup>7</sup>Department of Development and Regeneration, KU Leuven, Leuven, Belgium

<sup>8</sup>Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, Netherlands

<sup>9</sup>EPI-center, KU Leuven, Leuven, Belgium

## Corresponding author:

Alexander Pate, Jean McFarlane, University of Manchester, Manchester M13 9GB, UK.

Email: [alexander.pate@manchester.ac.uk](mailto:alexander.pate@manchester.ac.uk)

## I Introduction

Clinical prediction models (CPMs) are developed to predict expected health outcomes, such as an individual's probability that a specific disease or condition is present (diagnostic models) or that a specific event will occur in the future (prognostic models).<sup>1,2</sup> Logistic regression is typically used for developing CPMs to predict a single binary outcome. Often though, healthcare outcomes have multiple levels (multi-category/ polytomous outcomes), such as cancer grade or Likert scales. Then, the natural extension is to use multinomial logistic regression to develop the CPM. Multinomial models have been used to develop CPMs across a range of clinical settings,<sup>3–8</sup> and it has been argued they should be used to develop prediction models more often.<sup>9</sup> It has also been shown that multinomial logistic regression is preferred over multiple binary logistic regression when predicting multiple correlated binary outcomes to estimate their joint probability.<sup>10</sup>

An important design aspect when developing any prediction model is ensuring the sample size of the development dataset is appropriate to minimise overfitting and ensure sufficiently precise predictions. Sample size guidance for developing prediction models with continuous, binary and time-to-event outcomes have recently been developed.<sup>11–15</sup> However, there is a paucity of guidance for multinomial prediction models. Work by de Jong et al.,<sup>16</sup> highlighted the importance of considering the number of events per predictor for each outcome category when choosing the sample size, and showed that multinomial logistic regression models were susceptible to overfitting when fit in development data of small-to-medium sample size. However, there is no evidence to support events per predictor rules-of-thumb for calculating the required sample size,<sup>13,17</sup> and more tailored guidance is required.

Therefore, the aim of this study was to extend the existing sample size criteria by Riley et al.,<sup>11,12</sup> to cater for multinomial logistic regression prediction models predicting nominal polytomous outcomes. The remainder of this paper is structured as follows: the 'Existing sample size proposal for developing prediction models using binary logistic regression' section briefly reviews the minimum sample size criterion outlined by Riley et al.<sup>12</sup> for binary CPMs, and the 'Extending the sample size formula to multinomial logistic regression' section uses these as the foundation for our proposed sample size criteria for developing a multinomial logistic regression model. A detailed description of the simulation used to verify one of the proposed sample size criteria is given in Appendix S1. The 'Practical recommendations for implementing criteria in practice (estimating  $R^2_{CS\_adj}$  and dealing with large required sample sizes)' and 'A worked example of calculating sample size criteria for a multinomial logistic regression model' sections illustrate and advise on how to implement the proposed criteria in practice. Finally, in the 'Discussion' section we summarise the findings.

## 2 Existing sample size proposal for developing prediction models using binary logistic regression

We use the sample size criteria proposed by Riley et al.<sup>12</sup> as the basis for our extensions into multinomial logistic regression. In this section, we introduce the notation required for our proposals, but refer readers to previous literature<sup>11,12</sup> for a full discussion.

Consider a binary outcome,  $Y_i$  ( $i = 1, \dots, N$ ), which takes the value 1 if observation  $i$  has the outcome and is 0 otherwise. CPMs for such outcomes aim to estimate the probability of  $Y_i = 1$  conditional on a set of  $Q$  (candidate) predictor parameters, denoted as  $X_{qi}$  for  $q = 1, \dots, Q$ , collectively in the vector  $X_i = (X_{1i}, \dots, X_{Qi})^T$ . Note that predictor parameters refer to the number of coefficients that must be estimated in the model, rather than the number of covariates included in the model. This can be modelled using logistic regression to estimate  $P(Y_i = 1|X_i)$ , as

$$\log\left(\frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)}\right) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_Q X_{Qi}, \quad (1)$$

where  $\beta_1, \dots, \beta_Q$  are a set of predictor coefficients (conditional log odds ratios), which are estimated through maximum likelihood estimation to give estimates  $\hat{\beta}_1, \dots, \hat{\beta}_Q$ .

The Riley et al. sample size criteria for developing a binary CPM based on equation (1) have three components detailed in Table 1:

**Table 1.** Three component for deriving a minimum sample size for a binary logistic regression model.

<b>Criterion (i):</b> targeting the global shrinkage factor to be above a pre-defined threshold
<b>Criterion (ii)</b> targeting a small absolute difference in the apparent and adjusted Nagelkerke's $R^2$ ( $R^2_{Nagelkerke}$ ) <sup>18</sup>
<b>Criterion (iii)</b> targeting a precise estimate of overall risk (model intercepts).

We explain these in more detail and introduce the necessary notation for the rest of the manuscript in this section, and extend each criterion to multinomial logistic regression in the “Extending the sample size formula to multinomial logistic regression” section.

### 2.1 Overview of criterion (i): Sample size to target the global shrinkage factor to be above a pre-defined threshold

The first sample size criterion of Riley et al.<sup>12</sup> assesses overfitting on the multiplicative scale by considering shrinkage of predictor effects. This is when regression coefficients are shrunk towards zero to help mitigate against risk of overfitting. Criterion (i) is based on a global shrinkage factor ( $S$ ) that is applied to all predictor effects. Specifically, one multiplies  $\hat{\beta}_1, \dots, \hat{\beta}_Q$  of equation (1) by  $S$ , giving,

$$\log\left(\frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)}\right) = \alpha^* + S(\hat{\beta}_1 X_{1i} + \dots + \hat{\beta}_Q X_{Qi}), \tag{2}$$

where  $\alpha^*$  is the revised intercept to ensure the mean predicted risk matches the mean observed risk.<sup>19</sup> For the sample size criteria,<sup>12</sup> the van Houwelingen and Le Cessie’s heuristic shrinkage factor ( $S_{VH}$ )<sup>20</sup> is used to estimate  $S$ :

$$S_{VH} = 1 - \frac{Q}{LR}, \tag{3}$$

where  $Q$  is the number of candidate predictors parameters considered for inclusion prior to any variable selection, and  $LR = -2(\ln L_{null} - \ln L_{model})$  is the likelihood ratio statistic.

Criterion (i) of Riley et al.,<sup>12</sup> calculates a sample size  $n$  to target the shrinkage ( $S_{VH}$ ) to be above a pre-specified threshold (commonly taken as 0.9 or above, to target an overfitting of 10% or less, which leads to greater model stability<sup>21,22</sup>). For binary logistic regression, the required sample size to target a shrinkage factor  $S_{VH}$ , is calculated as:

$$n = \frac{Q}{(S_{VH} - 1) \log\left(1 - \frac{R_{CS\_adj}^2}{S_{VH}}\right)}, \tag{4}$$

where  $R_{CS\_adj}^2$  is an optimism-adjusted estimate of the Cox-Snell<sup>23</sup>  $R_{CS}^2$ .

### 2.2 Overview of criterion (ii): Ensuring small absolute difference in the apparent and adjusted $R_{Nagelkerke}^2$

The second sample size criterion of Riley et al.<sup>12</sup> is defined to ensure a small difference ( $\delta$ ) between the apparent and adjusted Nagelkerke  $R^2$ . It requires pre-specifying a value for  $\delta$  that one would tolerate, with small values preferred to improve model stability.<sup>21,22</sup> For any generalised linear model, Nagelkerke  $R^2$  is expressed as

$$R_{Nagelkerke}^2 = \frac{R_{CS}^2}{\max(R_{CS}^2)}, \tag{6}$$

where  $R_{CS}^2$  could be the apparent or optimism-adjusted estimate of  $R_{CS}^2$ . The maximum value of  $R_{CS}^2$  is calculated as  $\max(R_{CS}^2) = \max(R_{CS\_app}^2) = \max(R_{CS\_adj}^2) = 1 - \exp\left(\frac{2 \ln L_{null}}{n}\right)$ , where  $\ln L_{null}$  is the log-likelihood of the intercept-only model. It then follows that

$$R_{Nagelkerke\_app}^2 - R_{Nagelkerke\_adj}^2 = \frac{R_{CS\_app}^2}{\max(R_{CS}^2)} - \frac{R_{CS\_adj}^2}{\max(R_{CS}^2)} \leq \delta \tag{6}$$

holds if the required level of shrinkage ( $S_{VH}$ ) is such that:

$$S_{VH} \geq \frac{R_{CS\_adj}^2}{R_{CS\_adj}^2 + \delta \max(R_{CS}^2)}. \tag{7}$$

For binary logistic regression, the sample size for criterion (ii) is then calculated by substituting the minimum  $S_{VH}$  that satisfies equation (7) into equation (4).

### 2.3 Overview of criterion (iii): Ensure precise estimate of overall risk

The third sample size criterion of Riley et al.<sup>12</sup> is to ensure a precise estimate of overall risk. For binary logistic regression, an approximate 95% confidence interval for the estimate of the overall outcome proportion ( $\hat{\theta}$ ) can be expressed as

$$\hat{\theta} \pm 1.96 \sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}. \quad (8)$$

Therefore, to target a pre-specified absolute margin of error of  $\delta$ , the following sample size is required:

$$n = \left(\frac{1.96}{\delta}\right)^2 \hat{\theta}(1 - \hat{\theta}). \quad (9)$$

The final (minimum) sample size is then taken to be the maximum sample size across criterion (i), (ii) and (iii).

## 3 Extending the sample size formula to multinomial logistic regression

In this section, we extend each of the criteria from the previous section to the situation where the outcome has multiple categories, and we wish to develop a CPM using multinomial logistic regression. As such, hereto we consider an outcome,  $Y_i$  ( $i = 1, \dots, N$ ), which has  $K$  nominal categories, where  $Y_i = k$  for  $k \in [1, K]$  if individual  $i$  has the  $k$ th outcome category. All other notation introduced in the ‘Existing sample size proposal for developing prediction models using binary logistic regression’ section remains the same.

### 3.1 Introducing the multinomial logistic regression model and its calibration framework

A multinomial logistic regression model<sup>24</sup> predicting outcome,  $Y_i$ , with  $K$  nominal categories (taking the first category as the reference, without loss of generality), and  $Q$  the number of candidate predictor parameters in each sub-model  $k$ , is expressed by the following set of equations (dropping the subscript  $i$  for brevity):

$$\begin{aligned} P(Y = 1) &= \frac{1}{1 + \exp(\beta_{0,2} + \sum_{q=1}^Q \beta_{q,2} X_q) + \dots + \exp(\beta_{0,K} + \sum_{q=1}^Q \beta_{q,K} X_q)} \\ P(Y = 2) &= \frac{\exp(\beta_{0,2} + \sum_{q=1}^Q \beta_{q,2} X_q)}{1 + \exp(\beta_{0,2} + \sum_{q=1}^Q \beta_{q,2} X_q) + \dots + \exp(\beta_{0,K} + \sum_{q=1}^Q \beta_{q,K} X_q)} \\ &\vdots \\ &\vdots \\ P(Y = K) &= \frac{\exp(\beta_{0,K} + \sum_{q=1}^Q \beta_{q,K} X_q)}{1 + \exp(\beta_{0,2} + \sum_{q=1}^Q \beta_{q,2} X_q) + \dots + \exp(\beta_{0,K} + \sum_{q=1}^Q \beta_{q,K} X_q)} \end{aligned} \quad (10)$$

which equates to the following  $K - 1$  submodels:

$$\ln \left[ \frac{P(Y = k)}{P(Y = 1)} \right] = \beta_{0,k} + \sum_{q=1}^Q \beta_{q,k} X_q \quad (11)$$

for  $k = 2, \dots, K$ , alongside the constraint  $\sum_{k=1}^K P(Y = k) = 1$ .

Sub-model specific shrinkage factors can be defined for a multinomial logistic regression based on the recalibration framework outlined by Van Hoorde et al.<sup>25</sup> Specifically, after fitting a multinomial logistic regression using maximum likelihood (equation (11)), a separate shrinkage factor  $S_{MN,k}$  is applied to all the  $\beta$ 's for each sub-model  $k$ , and each intercept

updated (to ensure calibration-in-the-large), as follows:

$$\ln \left[ \frac{P(Y = k)}{P(Y = 1)} \right] = \alpha_{0,k}^* + S_{MN,k} \left( \sum_{q=1}^Q \hat{\beta}_{q,k} X_q \right) \tag{12}$$

for  $k = 2, \dots, K$ , where  $\hat{\beta}_{m,k}$  are the maximum likelihood estimates from the multinomial logistic regression (equation (11)), and  $\alpha_{0,k}^*$  are the re-estimated intercepts.

### 3.2 Extending criterion (i) to multinomial logistic regression

#### Direct application of criterion (i) to multinomial logistic regression models

A natural starting point for criterion (i) from binary logistic regression to multinomial logistic regression, would be to again base the required sample on targeting the single heuristic shrinkage factor of van Houwelingen and Le Cessie<sup>20</sup> to be at or above the chosen threshold. If we let

$$S_{VH\_MN} = 1 - \frac{(K - 1) * Q}{LR_{MN}}, \tag{13}$$

be the heuristic shrinkage factor of the multinomial model, where  $LR_{MN} = -2(\ln L_{null} - \ln L_{model})$  is like likelihood ratio test statistic for the multinomial model,  $(K - 1) * Q$  is the total number of candidate predictor parameters across all sub-models, and  $R_{CS\_adj}^2 = 1 - \exp\left(\frac{-LR_{MN}}{n}\right)$  is the apparent estimate of the Cox-Snell generalised definition of  $R^2$  for the multinomial model, then equation (4) could again be used to define a minimum required sample size to target  $S_{VH\_MN}$  to be at a pre-specified threshold:

$$n = \frac{(K - 1) * Q}{(S_{VH\_MN} - 1) \ln \left( 1 - \frac{R_{CS\_adj}^2}{S_{VH\_MN}} \right)}. \tag{14}$$

However, this approach has issues. While the van Houwelingen and Le Cessie<sup>20</sup> heuristic shrinkage factor is an estimator of the (true)  $S$  for binary logistic regression, there is no clear relationship between  $S_{VH\_MN}$  and the  $K - 1$  multinomial sub-model specific shrinkage factors ( $S_{MN,k}: k = 2, \dots, K$ ) in equation (12). Therefore, simply ensuring that  $S_{VH\_MN}$  surpasses a pre-specified threshold (using equation (14)) would not necessarily result in the required level of shrinkage in each sub-model (i.e.  $S_{MN,k}: k = 2, \dots, K$  in equation (12)). This is undesirable because it would mean that some sub-models of the multinomial model could be overfit. Therefore, we propose another approach to extend criterion (i), which targets all sub-model shrinkage factors ( $S_{MN,k}: k = 2, \dots, K$ ) to be at or above the desired threshold.

#### Alternative suggestion for criterion (i), utilising distinct logistic regression models

Equation (11) can be expressed as a set of  $K - 1$  distinct logistic regression models fitted separately, in the subset of the cohort which has either outcome  $k$  or outcome 1 (the reference). That is, the following binary logistic regression model can be fitted,

$$\ln \left[ \frac{P(Y = k)}{P(Y = 1)} \right] = \gamma_{0,k} + \sum_{q=1}^Q \gamma_{q,k} X_q, \tag{15}$$

on the subset of individuals where  $Y \in \{1, k\}$ , separately for  $k = 2, 3, \dots, K$ . These are also referred to as ‘one vs one’ models.<sup>6</sup>

Crucially, a separate shrinkage factor for each distinct logistic regression model can then be calculated,<sup>26–28</sup> such that

$$\ln \left[ \frac{P(Y = k)}{P(Y = 1)} \right] = \gamma_{0,k}^* + S_{DL,k} \left( \sum_{q=1}^Q \hat{\gamma}_{q,k} X_q \right), \tag{16}$$

where  $\hat{\gamma}_{m,k}$  are the coefficients estimated from equation (15),  $\gamma_{0,k}^*$  are the re-estimated intercepts, and  $S_{DL,k}$  is the shrinkage factor for distinct logistic regression model  $k$ , defined in the same way as  $S$  from the ‘Overview of criterion (i): Sample size to target the global shrinkage factor to be above a pre-defined threshold’ section (now referred to as ‘distinct logistic shrinkage factors’). This means that if an estimate of  $R_{CS\_adj}^2$  is available for each distinct logistic regression model  $k$ , then using the process summarised in the ‘Overview of criterion (i): Sample size to target the global shrinkage factor to be above a pre-defined threshold’ section, equation (4) can be used to derive a sample size to target a particular distinct logistic

shrinkage factor for each model. Importantly, it has been shown that the sub-models of the multinomial logistic regression, and the distinct logistic regression models, are parametrically equivalent ( $\gamma_{m,k} = \beta_{m,k}$ ).<sup>29</sup> Given the asymptotically unbiased property of the maximum likelihood estimators, it follows that  $\hat{\gamma}_{m,k} = \hat{\beta}_{m,k}$  as  $N \rightarrow \infty$ , and hence  $S_{DL,k} \rightarrow S_{MN,k}$  as  $N \rightarrow \infty$ . Therefore, deriving a sample size to target the shrinkage factor of each distinct logistic regression ( $S_{DL,k}$ :  $k = 2, \dots, K$ ) is above the desired value using equation (4),<sup>12</sup> will also target the multinomial sub-model specific shrinkage factors ( $S_{MN,k}$ :  $k = 2, \dots, K$ ) to be above the desired threshold. A separate sample size calculation must therefore be done for each pair of outcomes, taking the maximum to ensure criterion (i) is satisfied for each sub-model. One further point, this strategy relies on the fact that  $S_{DL,k} \rightarrow S_{MN,k}$  as  $N \rightarrow \infty$ . For the sample size criteria to work, we need close agreement between  $S_{DL,k}$  and  $S_{MN,k}$  at the value of  $N$  that satisfies the sample size criteria. We therefore report the agreement between  $S_{DL,k}$  and  $S_{MN,k}$  in the simulation carried out in the supplementary material, at a sample size  $N$  that meets said criteria.

#### Consideration of the choice of reference category

So far, the first outcome category has been taken as the reference. During model development, changing the reference category will not have an impact on the risk scores generated from the model. However, upon validating a multinomial CPM, the choice of reference category will change which of the multinomial sub-model specific shrinkage factors are calculated (i.e. what is estimated from equation (12)). Therefore, for the purposes of criterion (i), one must ensure that the shrinkage of every possible sub-model across all reference categories at model validation is above a certain level. Not doing so (i.e. only focusing on one outcome category), would potentially create over-confidence in the model's ability to distinguish between some of the outcome categories. In other words, one must aim to minimise optimism in all pairwise performance metrics. While calculating criterion (i) with taking each outcome category as a reference in turn may lead to high required sample sizes (e.g. see the 'A worked example of calculating sample size criteria for a multinomial logistic regression model' section), this would be reflective of one trying to develop a CPM to predict multinomial outcomes that require a lot of statistical power; this should be viewed as valuable information rather than a hindrance (as with any sample size calculation). We, therefore, outline our final approach in the following section, to ensure overfitting is minimised across all pairs of outcomes.

#### Final proposal for criterion (i)

Let  $S_{MN,k,r}$  be the sub-model specific shrinkage factor from the van Hoorde et al.<sup>25</sup> framework, for sub-model  $k$  with the reference category  $r$ . The following approach will target every  $S_{MN,k,r}$  to be above the pre-specified threshold. The proposal is to follow the approach outlined in the 'Alternative suggestion for criterion (i), utilising distinct logistic regression models' section for every possible reference category, and take the maximum sample size across all reference categories. That is for each distinct logistic regression model  $\{k, r\}$  where  $k \neq r$ ,

$$\ln \left[ \frac{P(Y = k)}{P(Y = r)} \right] = \gamma_{0,k,r} + \sum_{q=1}^Q \gamma_{q,k,r} X_q, \quad (17)$$

we can obtain a corresponding shrinkage factor,  $S_{DL,k,r}$ , each defined in the same way as  $S_{DL,k}$  from equation (16); that is,

$$\ln \left[ \frac{P(Y = k)}{P(Y = r)} \right] = \gamma_{0,j,r}^* + S_{DL,k,r} \left( \sum_{q=1}^Q \hat{\gamma}_{q,k,r} X_q \right). \quad (18)$$

Define  $m_{k,r}$  as the number of individuals with outcome category  $k$  or  $r$  that is required to target the shrinkage factor  $S_{DL,k,r}$  to be above some pre-defined threshold (e.g. 0.9). Here,  $m_{k,r}$  can be calculated using the existing formula for binary logistic regression<sup>12</sup>; specifically, using equation (4). To do so, appropriate estimates of  $R_{CS,k,r}^2$  (Cox-Snell  $R^2$  for distinct logistic regression model  $\{k, r\}$ ),  $Q$  (the number of candidate predictor parameters considered for inclusion in each sub-model), and  $p_k$  (the proportion of individuals from the cohort expected to have outcome  $k$ ) must each be pre-specified. Suggestions of how to pre-specify  $R_{CS,k,r}^2$  are given in the 'Practical recommendations for implementing criteria in practice (estimating  $R_{CS-adj}^2$  and dealing with large required sample sizes)' section. Note that because the logistic regression models for  $\ln \left[ \frac{P(Y=k)}{P(Y=r)} \right]$  and  $\ln \left[ \frac{P(Y=r)}{P(Y=k)} \right]$  are equivalent, we can reduce this to only consider the combinations where  $k > r$ .

The total number of individuals required in the whole cohort ( $n_{k,r}$ ) to ensure there are  $m_{k,r}$  individuals with outcome categories  $\{k, r\}$ , can then be calculated as  $n_{k,r} = m_{k,r} / p_{k,r}$ , where  $p_{k,r}$  is the proportion of individuals from the whole

cohort expected to have outcome categories  $\{k, r\}$ . Finally, the required sample size  $n$  to satisfy our criterion (i) is taken to be  $n = \max(n_{k,r}; k > r)$ .

The proposed approach to implementing criterion (i) are evaluated in a simulation study with full details provided in Appendix S1.

### 3.3 Extending criterion (ii) to multinomial logistic regression

As noted in the ‘Overview of criterion (ii): Ensuring small absolute difference in the apparent and adjusted  $R^2_{\text{Nagelkerke}}$ ’ section, the second criterion outlined by Riley et al.,<sup>12</sup> is defined to ensure a small difference ( $\delta$ ) between the observed and expected proportion of variance explained ( $R^2_{\text{Nagelkerke}}$ ) for the overall model.

As outlined in de Jong et al.,<sup>16</sup> the apparent  $R^2_{\text{Nagelkerke}}$  for a multinomial logistic regression model is defined in the same as for a binary logistic model:

$$R^2_{\text{Nagelkerke}} = \frac{1 - \exp\left(\frac{-\text{LR}_{MN}}{n}\right)}{1 - \exp\left(\frac{2\ln L_{null}}{n}\right)} = \frac{R^2_{CS}}{\max(R^2_{CS\_app})} \quad (19)$$

where  $\text{LR}_{MN}$  is defined as previously, and  $\ln L_{null}$  is the log-likelihood of an intercept only multinomial model. Therefore, given the definition of  $S_{VH\_MN}$  in equation (13), to ensure a difference of less than  $\delta$  between the apparent and adjusted  $R^2_{\text{Nagelkerke}}$  the following equation must hold:

$$S_{VH\_MN} \geq \frac{R^2_{CS\_adj}}{R^2_{CS\_adj} + \delta \max(R^2_{CS\_app})} \quad (20)$$

Plugging this into equation (14), and noting that  $n$  is a monotonically increasing function of  $S_{VH\_MN}$ , we get the following requirement for criterion (ii) for multinomial logistic regression prediction models:

$$n \geq \frac{(K - 1) * Q}{\left(\frac{R^2_{CS\_adj}}{R^2_{CS\_adj} + \delta \max(R^2_{CS\_app})} - 1\right) \ln(1 - R^2_{CS\_adj} - \delta \max(R^2_{CS\_app}))} \quad (21)$$

Given  $R^2_{\text{Nagelkerke}}$  is similarly defined for binary logistic and multinomial logistic regression models, this criterion is directly transferable from binary logistic regression to multinomial models. In line with the criterion (the ‘Overview of criterion (ii): Ensuring small absolute difference in the apparent and adjusted  $R^2_{\text{Nagelkerke}}$ ’ section) for binary logistic regression, we recommend a difference of  $\delta = 0.05$ .<sup>11,12</sup>

We note that for criterion (ii), we focus on the fit of the overall multinomial logistic regression model, in contrast to criterion (i) where we focused on each sub-model. The reason for this is that  $R^2_{CS}$  (and hence  $R^2_{\text{Nagelkerke}}$ ) is not typically expressed for the sub-models of a multinomial logistic regression. While we could ensure that criterion (ii) holds for each distinct logistic regression model, it is not clear what this would achieve with respect to the sub-models of the multinomial logistic regression model.

### 3.4 Extending criterion (iii) to multinomial logistic regression

As outlined in the ‘Overview of criterion (iii): Ensure precise estimate of overall risk’ section, the third criterion of Riley et al.,<sup>12</sup> is to ensure a precise estimate of overall risk (i.e. model intercept). To mimic the approach for binary logistic regression, for a multinomial model, this can be approximated by calculating the margin of error in the outcome proportion estimates.

Let  $p_k = E_k / n$  be the proportion of individuals from the entire cohort with outcome category  $\{k\}$ , with  $E_k$  the number of events in outcome category  $k$ . If  $\pi_k$  is the underlying multinomial probability of outcome category  $k$ , then it can be shown through the work of Quesenberry and Hurst,<sup>30</sup> and Goodman,<sup>31</sup> that the simultaneous  $\alpha \times 100\%$  confidence interval limits for  $\pi_1, \pi_2, \dots, \pi_K$ :

$$\pi_k^- \leq \pi_k \leq \pi_k^+; k = 1, \dots, K$$

can be estimated by

$$p_k \pm \left[ \chi_{\frac{\alpha}{2}, 1}^2 \times \frac{p_k(1-p_k)}{n} \right]^{\frac{1}{2}} \quad (22)$$

where  $\chi_{\frac{\alpha}{2}, 1}^2$  denotes the Chi-squared distribution with 1 degree of freedom. Therefore, the sample size to ensure an absolute margin of error  $\delta$  (say 0.05) at a  $(1 - \alpha) \times 100\%$  confidence level is

$$n = \max_{k=1, \dots, K} \left( \frac{\chi_{\frac{\alpha}{2}, 1}^2 \times p_k(1-p_k)}{\delta^2} \right). \quad (23)$$

We choose to target simultaneous confidence intervals<sup>30,31</sup> rather than pointwise confidence intervals so that every estimate of overall risk will simultaneously be within the pre-defined margin of error. This will require a larger sample size than considering pointwise confidence intervals and is therefore conservative.

It is important to mention that we are primarily interested in a precise estimate of the mean risk of each outcome category across all individuals in the population after adjustment for predictors. However, the mean risk of each outcome category across all individuals will often be similar to the outcome proportions observed from a null model with no predictors (which we are working with above). The variability of these two quantities will therefore also be similar, and we can approximate the variability of the mean risk of each outcome category in the population using the above formula.

A summary of our proposed sample size criteria for a multinomial logistic regression model is given in Table 2.

#### 4 Practical recommendations for implementing criteria in practice (estimating $R_{CS\_adj}^2$ and dealing with large required sample sizes)

To perform our proposed sample size calculations, an estimate of  $R_{CS\_adj}^2$  needs to be pre-specified. As with earlier work<sup>11,12,14</sup> we recommend that this is based on similar, previously developed or validated prediction models. When calculating criterion (i), estimates of  $R_{CS\_adj, k, r}^2$  are required for each distinct logistic regression model  $\{k, r\}$ , corresponding to

**Table 2.** Summary of the proposed minimum sample size criteria for multinomial logistic regression CPMs

##### Step 1: Choose number of predictor parameters $Q$ considered for inclusion in each sub-model at model development

Recognise that one predictor may require  $> 1$  predictor parameter; for example, categorical predictor  $r$  with  $> 2$  categories, a continuous predictor with nonlinear terms, and interaction terms.

##### Step 2: Choose sensible values for $p_k$ and $p_{k,r}$ , the proportion of individuals in the cohort with outcomes in category $k$ and $\{k, r\}$ $\max(R_{CS\_app}^2)$ and $R_{CS\_adj}^2$ for the multinomial model, and $R_{CS\_adj, k, r}^2$ of each distinct logistic regression model

Ideally, this will be based on previously published models in the same setting with similar outcome definition, a variety of ways to estimate these from various reported statistics are given in the 'Practical recommendations for implementing criteria in practice (estimating  $R_{CS\_adj, k, r}^2$  and dealing with large required sample sizes)' section. If no previous information is available to estimate  $R_{CS\_adj, k, r}^2$ , use values which correspond to an  $R_{Nagelkerke}^2 = 0.15$  in each sub-model.

##### Step 3: Criterion (i)

1. Calculate the minimum sample size ( $m_{k,r}$ ) for each distinct logistic regression model  $\{k, r\}$ , where  $k > r$ , using equation (4) based on a pre-specified level of shrinkage (for example, targeting shrinkage factors of 0.9) and an estimate of  $R_{CS\_adj, k, r}^2$ .
2. Calculate the total number of individuals needed to achieve the required number in each distinct logistic regression model  $\{k, r\}$ , by dividing by  $p_{k,r}$ ,  $n_{k,r} = m_{k,r} / p_{k,r}$
3. Take the minimum sample size for criterion (i) to be  $n = \max(n_{k,r}; k > r)$ , which will target all the multinomial sub-model specific shrinkage factors to be greater-than-or-equal to the pre-specified threshold.

##### Step 4: Criterion (ii)

Use equation (21) to calculate a sample size to target the difference between the apparent and optimism adjusted  $R_{Nagelkerke}^2$  to be  $\delta$ , using estimates of  $\max(R_{CS\_app}^2)$  and  $R_{CS\_adj}^2$ . Previously  $\delta = 0.05$  has been recommended.<sup>14</sup>

##### Step 5: Criterion (iii)

Use equation (23) to calculate a sample size to target the simultaneous 95% confidence intervals of the estimates of overall risk for each category to be  $\leq \delta$ , using estimates of  $p_k$ . We recommend  $\delta = 0.05$ .

##### Step 6: Final sample size

The required minimum sample size is the maximum value from steps 3 to 5, to ensure criteria (i), (ii) and (iii) are met.



the sub-models of the multinomial logistic model. When calculating criterion (ii) an estimate of  $R^2_{CS\_adj}$  is required for the multinomial logistic regression model. We discuss how to estimate these using the published data below. We also urge researchers to report the metrics discussed below when publishing future CPM development papers based on multinomial logistic regressions, to aid the sample size calculations of others. Finally, we make recommendations on what to do if the calculated required sample size is unfeasibly high.

#### 4.1 Recommendations for deriving $R^2_{CS\_adj,k,i}$ of distinct logistic regression models

To calculate criterion (i) estimates of  $R^2_{CS\_adj,k,r}$  are required. If the appropriate ‘one-vs-one’<sup>6</sup> distinct logistic regression models have been fitted in a published study and estimates of  $R^2_{CS\_adj,k,r}$  have been reported, these can be used directly. If other pseudo-  $R^2$  statistics have been reported (for each distinct logistic), there are a variety of ways to derive  $R^2_{CS\_adj,k,r}$  from these; see Riley et al.<sup>12</sup> Alternatively, if the C-statistics of each distinct logistic regression model are available, then  $R^2_{CS\_adj,k,r}$  can be estimated using a simulation approach.<sup>15</sup>

However, it is highly likely that each distinct logistic regressions will not have been fitted alongside any previously developed multinomial logistic regression model. In this case, the pairwise C-statistics<sup>32</sup> (using the conditional risk method) of the multinomial logistic regression might have been reported. Here, since these pairwise C-statistics provide an estimate of the C-statistic for each distinct logistic regression model, they can be used to estimate  $R^2_{CS\_adj,k,r}$  using the simulation approach of Riley et al.<sup>15</sup> We illustrate this approach in our worked example in the ‘A worked example of calculating sample size criteria for a multinomial logistic regression model’ section.

If neither pseudo-  $R^2$  or (pairwise) C-statistics are available a priori, we suggest calculating the minimum sample size following the approach suggested by Riley et al.,<sup>14</sup> for when information on  $R^2_{CS\_adj}$  is not available. Specifically, under a conservative assumption of optimism adjusted  $R^2_{Nagelkerke}$  of 0.15 (15%), equation (5) can be modified to give  $R^2_{CS\_app,k,r} = 0.15 * \max(R^2_{CS\_app,k,r})$  for each distinct logistic regression model. Here,  $\max(R^2_{CS\_app,k,r})$  can be estimated for each model using equation (6):

$$\max(R^2_{CS\_app,k,r}) = 1 - \exp\left(\frac{2 * \ln L_{null,k,r}}{n}\right), \tag{24}$$

where  $\ln L_{null,k,r}$  can be calculated for each distinct logistic regression model using:

$$\ln L_{null,k,r} = E_k \ln\left(\frac{E_k}{E_k + E_r}\right) + E_r \ln\left(\frac{E_r}{E_k + E_r}\right), \tag{25}$$

where  $E_k$  and  $E_r$  are the number of outcome events in the category  $k$  and  $r$ , respectively. Alternatively (and equally), for each distinct logistic regression model  $\max(R^2_{CS\_app,k,r})$  can be calculated as:

$$\max(R^2_{CS\_app,k,r}) = 1 - (\varphi_{k,r}^{\varphi_{k,r}} (1 - \varphi_{k,r})^{1-\varphi_{k,r}})^2, \tag{26}$$

where  $\varphi_{k,r} = E_k / E_k + E_r$ , is the outcome proportion in the category  $k$  relative to the reference category  $r$ . If a multinomial model had been published, then this information would be available for each distinct logistic regression model assuming the number of events in each category had been reported.

#### 4.2 Recommendations for deriving $R^2_{CS\_adj}$ of multinomial logistic regression models

To calculate criterion (ii) a pre-specified estimate of the overall  $R^2_{CS\_adj}$  is required. As previous, this would ideally be based on information from a previous multinomial logistic regression model. Similarly to binary logistic regression, if other pseudo- $R^2$  statistics have been reported, there are a variety of ways to derive  $R^2_{CS\_adj}$  from these, as outlined in Riley et al.<sup>12</sup>

Alternatively, one could again take a conservative approach to setting  $R^2_{CS\_adj} = 0.15 * \max(R^2_{CS\_app})$  (corresponding to an  $R^2_{Nagelkerke}$  of 0.15). There are two ways to calculate  $\max(R^2_{CS\_app})$  for multinomial logistic regression. The first is to use equation (6), where  $\ln L_{null}$  can be calculated for a multinomial logistic regression as:

$$\ln L_{null} = \sum_{k=1}^K E_k \ln\left(\frac{E_k}{n}\right), \tag{27}$$

with  $E_k$  denoting the number of events in outcome category  $k$ . Alternatively,  $\max(R_{CS\_app}^2)$  can be expressed as:

$$\max(R_{CS\_app}^2) = 1 - \left( \prod_{k=1}^K (p_k)^{p_k} \right)^2, \quad (28)$$

where  $p_k = E_k / n$  is the observed frequency of category  $k$ , as defined in the ‘Extending criterion (iii) to multinomial logistic regression’ section. This expression follows naturally from equations (6) and (27), and details of its derivation are given in Appendix S1. Some implications of basing the estimate of  $R_{CS\_app}^2$  on the assumption  $R_{Nagelkerke}^2 = 0.15$  are also given in Appendix S1.

### 4.3 Recommendations if required sample size is too high

We propose three strategies if the required sample size is completely unfeasible to recruit. It is worth reiterating, that the estimated sample size is required to build the proposed model with the specified levels of overfitting, optimism and precision. In order to reduce the sample size, the model must either be simplified, or you must be willing to accept overfitting, optimism and precision below the desired level.

1. Merge outcome categories. We believe the first consideration could be to merge some of the outcome categories that are driving the high sample size; looking at each pairwise criterion (i) will indicate which categories are driving the sample size. This should only be done if it makes sense from a clinical point of view, and knowing the risks of the merged categories would be of clinical interest.
2. A second suggestion is to reduce the number of candidate predictor parameters considered for inclusion in the model, which is inline with previous suggestions.<sup>11,12,14</sup> However, we have only looked at scenarios where there are a fixed number of predictor parameters considered for each sub-model. This means when reducing the number of predictor parameters, one would be doing so across all sub-models. An alternative to this is to only reduce the number of predictor parameters considered for inclusion in the sub-model(s) with the highest level of overfitting. This is an enticing approach, as one does not want to reduce the number of predictor parameters in sub-models that are not suffering from overfitting. However, the implications of such an approach are not yet clear and would require further research before this could be recommended.
3. Reduce the acceptable level of overfitting between specific pairs of outcomes. Rather than having the acceptable level of shrinkage at 0.9, it could be reduced (e.g. to 0.8), specifically for the pair of outcomes that are driving the high sample size. This is somewhat undesirable as criterion (i) is in place to minimise overfitting. However, at least the targeted level of overfitting would be explicitly stated and the limitations of the model would therefore be well quantified.

## 5 A worked example of calculating sample size criteria for a multinomial logistic regression model

### 5.1 Hypothetical scenario and information available in literature

In this section, we present a worked example to illustrate how our proposed sample size criteria could be implemented in practice. The code that was used to do this is available on GitHub.<sup>33</sup> Our example aims to calculate the minimum sample size required to develop a multinomial logistic regression prediction model to predict the tumour type (benign, borderline, stage I invasive, stage II-IV invasive, or metastatic) when presented with an ovarian mass. This is an important preoperative diagnosis, as dependent on the type of tumour, different clinical action may be taken.

Van Calster et al.<sup>8</sup> considered the development of such a model using the International Ovarian Tumor Analysis Group<sup>34</sup> dataset. The following information is available from that work. The model was developed on a dataset of 3506 tumours, of which 2557 were benign, 186 were borderline, 176 were stage I invasive, 467 were stage II-IV invasive, and 120 were metastatic. The following pairwise C-statistics<sup>32</sup> were reported for every combination of outcome comparisons: 0.85 (benign vs borderline), 0.92 (benign vs stage I invasive), 0.99 (benign vs stage II-IV invasive), and 0.95 (benign vs metastatic), 0.75 (borderline vs stage I invasive), 0.95 (borderline vs stage II-IV invasive), 0.87 (borderline vs metastatic), 0.87 (stage I invasive vs stage II-IV invasive), 0.71 (stage I invasive vs metastatic) and 0.82 (stage II-IV invasive vs metastatic). These pairwise C-statistics are reported from a temporal validation and are free from in-sample optimism concerns, therefore we can use these to estimate  $R_{CS\_adj,k,i}^2$  directly with no adjustment for

optimism required. There were 17 candidate predictor parameters considered for inclusion in the model including all the fractional polynomials of continuous variables (each extra fractional polynomial term counts as an additional predictor parameter). We will assume we will consider the same set of variables for inclusion before applying variable selection techniques. We now illustrate the use of the aforementioned information to perform our sample size calculation.

## 5.2 Steps 1 and 2: Identifying values for $p_k$ , $p_{k,r}$ , $\max(R_{CS\_app}^2)$ , $R_{CS\_adj}^2$ , and $R_{CS\_adj,k,r}^2$

$k$  and  $r$  take the values 1 (benign), 2 (borderline), 3 (stage I invasive), 4 (stage II–IV invasive), and 5 (metastatic).

### Calculating $Q$

Assuming we consider the same set of variables for variable selection that were used in the work by Van Calster et al.,<sup>8</sup> this would mean  $Q = 17$ .

### Calculating $p_k$ and $p_{k,r}$

$p_k$  is the proportion of individuals that have outcome category  $\in \{k\}$ ,  $p_{k,r}$  is the proportion of individuals that have outcome category  $\in \{k, r\}$  where  $k > r$ . To estimate these values, we use the prevalence of each outcome category as reported in the ‘Hypothetical scenario and information available in literature’ section:  $p_1 = 0.729$ ,  $p_2 = 0.053$ ,  $p_3 = 0.050$ ,  $p_4 = 0.133$ ,  $p_5 = 0.034$ ,  $p_{2,1} = 0.782$ ,  $p_{3,1} = 0.780$ ,  $p_{4,1} = 0.863$ ,  $p_{5,1} = 0.764$ ,  $p_{3,2} = 0.103$ ,  $p_{4,2} = 0.186$ ,  $p_{5,2} = 0.087$ ,  $p_{4,3} = 0.183$ ,  $p_{5,3} = 0.084$ ,  $p_{5,4} = 0.167$ .

### Calculating $\max(R_{CS\_app}^2)$

We calculated  $\max(R_{CS\_app}^2)$  using equation (28), and the prevalence of each outcome category  $p_k$ :

$$\max(R_{CS\_app}^2) = 1 - (0.729^{0.729} * 0.053^{0.053} * 0.050^{0.050} * 0.133^{0.133} * 0.034^{0.034})^2 = 0.841.$$

### Calculating $R_{CS\_adj}^2$

Given the  $R_{CS\_adj}^2$  of the overall multinomial model had not been reported, we based our estimate of  $R_{CS\_adj}^2$  on assuming  $R_{Nagelkerke}^2 = 0.15$ . Using the estimate of  $\max(R_{CS\_app}^2)$  in equation (5) gave an estimate of:

$$R_{CS\_adj}^2 = 0.15 * \max(R_{CS\_app}^2) = 0.15 * 0.841 = 0.126.$$

### Calculating $R_{CS\_adj,k,r}^2$

No data was available on the  $R_{CS\_adj,k}^2$  in the work of Van Calster et al.,<sup>8</sup> therefore we had to estimate them indirectly using the simulation approach of Riley et al.<sup>15</sup> This method utilises the pairwise C-statistics, on which data was available.<sup>8</sup> Estimates of the pairwise outcome proportions  $\varphi_{k,r}$  of category  $k$  relative to the reference category  $r$  are also required to implement the simulation approach. These were estimated using the number of tumours in each category:  $\varphi_{2,1} = 0.068$ ,  $\varphi_{3,1} = 0.064$ ,  $\varphi_{4,1} = 0.154$ ,  $\varphi_{5,1} = 0.045$ ,  $\varphi_{3,2} = 0.486$ ,  $\varphi_{4,2} = 0.715$ ,  $\varphi_{5,2} = 0.392$ ,  $\varphi_{4,3} = 0.726$ ,  $\varphi_{5,3} = 0.405$ ,  $\varphi_{5,4} = 0.204$ . The simulation approach was followed for each sub-model to give estimates of:  $R_{CS\_adj,2,1}^2 = 0.116$ ,  $R_{CS\_adj,3,1}^2 = 0.179$ ,  $R_{CS\_adj,4,1}^2 = 0.497$ ,  $R_{CS\_adj,5,1}^2 = 0.170$ ,  $R_{CS\_adj,3,2}^2 = 0.185$ ,  $R_{CS\_adj,4,2}^2 = 0.499$ ,  $R_{CS\_adj,5,2}^2 = 0.374$ ,  $R_{CS\_adj,4,3}^2 = 0.328$ ,  $R_{CS\_adj,5,3}^2 = 0.129$ ,  $R_{CS\_adj,5,4}^2 = 0.210$ . Code for this simulation approach is provided at the GitHub repository,<sup>33</sup> as well as more general code on how to implement this simulation approach in the original work.<sup>15</sup>

## 5.3 Step 3: Criterion (i)

Following the process in the ‘Final proposal for criterion (i)’ section, first, each  $m_{k,r}$  was calculated using equation (4) and the estimates of  $R_{CS\_adj,k,r}^2$  from the ‘Calculating  $R_{CS\_adj,k,r}^2$ ’ section. Then the total number of individuals required to target a

multinomial sub-model specific shrinkage factor of 0.9 for sub-model  $\{k, r\}$ ,  $n_{k,r}$ , was calculated by dividing  $m_{k,r}$  by  $p_{k,r}$ :

$$\begin{aligned}
 n_{2,1} = m_{2,1} / p_{2,1} &= \frac{17}{(0.9 - 1) \ln\left(1 - \frac{0.116}{0.9}\right)} * \frac{3506}{2557 + 186} = 1574 \\
 n_{3,1} = m_{3,1} / p_{3,1} &= \frac{17}{(0.9 - 1) \ln\left(1 - \frac{0.179}{0.9}\right)} * \frac{3506}{2557 + 176} = 982 \\
 n_{4,1} = m_{4,1} / p_{4,1} &= \frac{17}{(0.9 - 1) \ln\left(1 - \frac{0.497}{0.9}\right)} * \frac{3506}{2557 + 467} = 246 \\
 n_{5,1} = m_{5,1} / p_{5,1} &= \frac{17}{(0.9 - 1) \ln\left(1 - \frac{0.170}{0.9}\right)} * \frac{3506}{2557 + 120} = 1067 \\
 n_{3,2} = m_{3,2} / p_{3,2} &= \frac{17}{(0.9 - 1) \ln\left(1 - \frac{0.185}{0.9}\right)} * \frac{3506}{186 + 176} = 7147 \\
 n_{4,2} = m_{4,2} / p_{4,2} &= \frac{17}{(0.9 - 1) \ln\left(1 - \frac{0.499}{0.9}\right)} * \frac{3506}{186 + 467} = 1128 \\
 n_{5,2} = m_{5,2} / p_{5,2} &= \frac{17}{(0.9 - 1) \ln\left(1 - \frac{0.374}{0.9}\right)} * \frac{3506}{186 + 120} = 3629 \\
 n_{4,3} = m_{4,3} / p_{4,3} &= \frac{17}{(0.9 - 1) \ln\left(1 - \frac{0.328}{0.9}\right)} * \frac{3506}{176 + 467} = 2045 \\
 n_{5,3} = m_{5,3} / p_{5,3} &= \frac{17}{(0.9 - 1) \ln\left(1 - \frac{0.129}{0.9}\right)} * \frac{3506}{176 + 120} = 13063 \\
 n_{5,4} = m_{5,4} / p_{5,4} &= \frac{17}{(0.9 - 1) \ln\left(1 - \frac{0.210}{0.9}\right)} * \frac{3506}{467 + 120} = 3813
 \end{aligned}$$

The minimum required sample size was taken as the maximum of these, and therefore  $N = 13063$ , approximately 9527 benign tumours, 693 borderline, 656 stage I invasive, 1740 stage II–IV invasive and 447 metastatic (assuming same outcome proportions as in Van Calster et al.<sup>8</sup>).

#### 5.4 Step 4: Criterion (ii)

Criterion (ii) aims to calculate a sample size required to ensure a difference of 0.05 between the apparent and adjusted  $R^2_{\text{Nagelkerke}}$ , which holds if equation (21) is satisfied ('Extending criterion (ii) to multinomial logistic regression' section). Plugging in the estimates of  $\max(R^2_{CS\_app})$  and  $R^2_{CS\_adj}$  into equation (21) gives:

$$n \geq \frac{4 * 17}{\left(\frac{0.126}{0.126 + 0.05 * 0.841} - 1\right) \ln(1 - 0.126 - 0.05 * 0.841)}$$

$$n \geq 1477.$$

So 1477 individuals are required to meet criterion (ii), approximately 1077 benign tumours, 78 borderline, 74 stage I invasive, 197 stage II–IV invasive, and 51 metastatic.

### 5.5 Step 5: Criterion (iii)

Criterion (iii) is to ensure a precise estimate of risk in the overall population. Following the steps outlined in the ‘Extending criterion (iii) to multinomial logistic regression’ section, for a 95% confidence interval ( $\alpha = 0.05$ ), using the estimated values for  $p_k$ , with  $K = 5$  and an absolute margin of error of  $\delta = 0.05$ , then the required sample size for each outcome is (equation (23)):

$$n_1 = \frac{6.635 \times 0.729(1 - 0.729)}{0.05^2} = 524$$

$$n_2 = \frac{6.635 \times 0.053(1 - 0.053)}{0.05^2} = 134$$

$$n_3 = \frac{6.635 \times 0.50(1 - 0.150)}{0.05^2} = 127$$

$$n_4 = \frac{6.635 \times 0.133(1 - 0.133)}{0.05^2} = 307$$

$$n_5 = \frac{6.635 \times 0.034(1 - 0.034)}{0.05^2} = 88$$

Leaving the sample size for criterion (iii) to be  $\max(n_1, n_2, n_3, n_4, n_5) = 524$ , approximately 382 benign tumours, 28 borderline, 26 stage I invasive, 70 stage II–IV invasive, and 18 metastatic.

### 5.6 Step 6: Final sample size

The final sample size is taken as the maximum of the sample sizes required to satisfy each criterion (i)–(iii), which was 13,063 (i), 1476 (ii) and 524 (iii) respectively. Hence, the minimum required sample size would be  $n = 13063$ , approximately 9527 benign tumours, 693 borderline, 656 stage I invasive, 1740 stage II–IV invasive and 447 metastatic (assuming same outcome proportions as in Van Calster et al.<sup>8</sup>). For contrast, using the definition of events per variable ( $EPV_m$ ) from De Jong et al.,<sup>16</sup> an  $EPV_m$  of 10 would result in a sample size of 4967, and an  $EPV_m$  of 20 would result in a sample size of 9934.

### 5.7 Suggestions for dealing with high sample size

The required sample size is high and is being driven by outcome categories 3 (stage I invasive) and 5 (metastatic). If the proposed model was developed with a sample size smaller than 13,063, the level of overfitting between these two outcomes would not be targeted at the pre-specified level of 0.9. Following the suggestions in the ‘Recommendations if required sample size is too high’ section, the first solution would be to merge categories 3 and 4 (stage I invasive with stage II–IV invasive). With such a combination the model would retain clinical interpretation. If it was essential to keep these outcome categories separate, fewer predictor parameters could be considered instead. The value of  $Q = 17$  incorporates fractional polynomial terms and interactions which could be removed, or one of the predictors could be removed altogether. A final possible option is to reduce the targeted level of overfitting for pair  $\{3, 5\}$ . Plugging a value of 0.8 into the ‘Step 3: Criterion (i)’ section would give  $n_{5,3} = 5746$ , and the final sample size would be driven by  $n_{3,2} = 7147$ . While this is slightly undesirable, the targeted level of overfitting for all other outcome pairs would still be at 0.9, and one could report that overfitting may be more likely for outcome pair  $\{3, 5\}$ .

## 6 Discussion

We have presented sample size criteria for the development of prediction models for multiple-category outcomes using multinomial logistic regression. This builds upon recent developments in this space for continuous, binary and time-to-event outcomes.<sup>11,12</sup> Criterion (ii) and (iii) both had a natural extension into a multinomial framework. Criterion (i) did not and therefore we tested the properties of our proposed approach in a simulation (Appendix S1), finding that the sample size resulted in the desired level of overfitting. Our approach to criterion (i) may lead to high sample sizes if some of the outcome categories are rare, or have a low pairwise  $R_{CS}^2$ , however this is necessary if you want to ensure overfitting is minimised in prediction between all pairs of outcome categories. If the required sample

size cannot be achieved, we have made some recommendations on how the model could be adjusted to lower the number required.

The biggest practical challenge with implementing these recommendations in practice is the availability of information on past  $R_{CS}^2$ , given multinomial logistic regression CPMs are not (yet) very common. The proposed criteria will be most effective in achieving their aim when an accurate estimate of the  $R_{CS}^2$  is available for both the multinomial model ( $R_{CS\_adj}^2$ ) and each distinct logistic regression model  $\{k, r\}$  ( $R_{CS\_adj,k,r}^2$ ). We have given advice on how to pre-specify these, but also want to urge researchers to report the relevant information when developing a multinomial logistic regression to enable this process. Currently, there is no way to estimate  $R_{CS\_adj}^2$  for the multinomial model from metrics which are not pseudo- $R^2$  (for example there is no way to estimate it from the PDI<sup>35</sup>), meaning reporting  $R_{CS\_adj}^2$  is very important. Estimates of  $R_{CS\_adj,k,r}^2$  can be obtained from a variety of metrics from previously published ‘one-to-one’ distinct logistic regression models.<sup>12,15</sup> However, when fitting a multinomial logistic regression, it is important to (at least) report the pairwise C-statistics using the conditional risk method<sup>32</sup> when fitting a multinomial logistic regression model. This is an informative performance metric that should be reported anyway, and it will allow future researchers to estimate  $R_{CS\_adj,k,r}^2$  (as was done in our worked example). In theory, the conditional risk method could also be used to report  $R_{CS\_adj,k,r}^2$  directly, although we are not aware of any instances of people doing this in the literature.

There are five important areas of future work. First, to establish a relationship between the PDI,<sup>35</sup> a commonly reported statistic for discrimination of a multinomial logistic regression model, and the distribution of the linear predictors of the sub-models. This relationship has been established for the C-statistic and logistic regression<sup>36–39</sup> allowing  $R_{CS}^2$  to be estimated when only the C-statistic is available.<sup>15</sup> Secondly, the simulation (Appendix S1) found that there was poor agreement between the heuristic shrinkage factors and the sub-model-specific shrinkage factors when covariate effects were large and sample sizes were small (Table S3). This resulted in not having the desired level of shrinkage in the developed models. This finding extends to binary logistic regression, but it is not clear whether similar results would be found for continuous or time-to-event outcomes. Given the proposed criterion (i) for every outcome type<sup>11,12</sup> targets the heuristic shrinkage factor to be at the chosen threshold, it’s important to establish in which scenarios where this may be a poor predictor of the sub-model specific shrinkage factors. Third, to extend the criteria of van Smeden et al.,<sup>13</sup> to multinomial logistic regression. Their work acts as a fourth criterion,<sup>14</sup> to target the mean absolute prediction error (MAPE) of a binary logistic regression model to be below a pre-specified threshold. This helps ensure precise predictions across the spectrum of predicted values. The formula is derived from a detailed simulation, in which a variety of binary logistic regression models are simulated and the MAPE assessed when the model is applied to new individuals from the target population. The first step to extending this criteria to multinomial logistic regression would be to define an extension of the MAPE for multinomial outcomes, which would then need to be followed by a similar simulation as the one used to derive the formula for binary logistic regression. The fourth is to develop sample size formula for the prediction of ordinal outcomes. The sample size formula proposed in this study are for a multinomial logistic regression, which can be fit to either nominal or ordinal outcomes. However if wanting to predict an ordinal outcome, an ordinal model could be fitted which would likely require a smaller sample size since they require less parameters to be estimated (for example if a proportional odds assumption is made). While the sample size criteria proposed in this paper would be valid for the prediction of an ordinal outcome using multinomial logistic regression, it is a conservative estimate. Future work could develop less conservative sample size criteria developed specifically for ordinal regression modelling techniques. In clinical trials for ordinal outcomes the proportional odds assumption has been shown to impact the required sample size,<sup>40</sup> and this would be no different for CPMs. The fifth is to explore the idea of reducing the number of candidate predictor parameters in specific sub-models of the multinomial logistic regression as a way to reduce the required sample size, as discussed in the ‘Recommendations if required sample size is too high’ section.

These sample size criteria will be embedded into existing software (pmsampsize in R<sup>41</sup> and Stata<sup>42</sup>) so they can be widely implemented in practice.

## Acknowledgements

The authors wish to thank anonymous reviewers and the Editors for their constructive comments which helped improve the article upon revision.

## Author contributions

AP, GM and RR conceived and designed the study. AP conducted the analysis and interpreted the results in discussion with GM and RR. AP wrote the initial draft of the manuscript with support from GM, which was then critically reviewed for important intellectual content by GM, RR, GSC, MVS, JE and BVC. JE will embed the methods into R and Stata software. All authors have approved the final version of the paper.

## Availability of data and materials

Full reusable code for running the simulation and worked example are provided at the referenced GitHub repository.<sup>33</sup>


## Declaration of conflicting interests


The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


## Funding


The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by funding from the MRC-NIHR Methodology Research Programme [grant number: MR/T025085/1]. GSC was supported by the NIHR Biomedical Research Centre, Oxford, and Cancer Research UK (programme grant: C49297/A27294).


## ORCID iDs

Alexander Pate  <https://orcid.org/0000-0002-0849-3458>

Gary S Collins  <https://orcid.org/0000-0002-2772-2316>

Ben Van Calster  <https://orcid.org/0000-0003-1613-7450>

Joie Ensor  <https://orcid.org/0000-0001-7481-0282>

Glen P Martin  <https://orcid.org/0000-0002-3410-9472>

## Supplemental material

Supplemental material for this article is available online.

## References

1. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015; **162**: 55–63.
2. van Smeden M, Reitsma JB, Riley RD, et al. Clinical prediction models: diagnosis versus prognosis. *J Clin Epidemiol* 2021; **132**: 142–145.
3. Wijesinha A, Begg CB, Funkenstein HH, et al. Methodology for the differential diagnosis of a complex data set. *Med Decis Mak* 1983; **3**: 133–154.
4. Schuit E, Kwee A, Westerhuis M, et al. A clinical prediction model to assess the risk of operative delivery. *BJOG An Int J Obstet Gynaecol* 2012; **119**: 915–923.
5. Barnes DE, Mehta KM, Boscardin WJ, et al. Prediction of recovery, dependence or death in elders who become disabled during hospitalization. *J Gen Intern Med* 2013; **28**: 261–268.
6. Van Calster B, Valentin L, Van Holsbeke C, et al. Polytomous diagnosis of ovarian tumors as benign, borderline, primary invasive or metastatic: development and validation of standard and kernel-based risk prediction models. *BMC Med Res Methodol* 2010; **10**: 96.
7. Roukema J, Van Leonhout RB, Steyerberg EW, et al. Polytomous regression did not outperform dichotomous logistic regression in diagnosing serious bacterial infections in febrile children. *J Clin Epidemiol* 2008; **61**: 135–141.
8. Van Calster B, Van Hoorde K, Valentin L, et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: prospective multi-centre diagnostic study. *Br Med J* 2014; **349**: 1–14.
9. Biesheuvel CJ, Vergouwe Y, Steyerberg EW, et al. Polytomous logistic regression analysis could be applied more often in diagnostic research. *J Clin Epidemiol* 2008; **61**: 125–134.
10. Martin GP, Riley RD, Sperrin M, et al. Clinical prediction models to predict the risk of multiple binary outcomes : a comparison of approaches. *Stat Med* 2020; **40**: 498–517.
11. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part I – continuous outcomes. *Stat Med* 2019; **38**: 1262–1275.
12. 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* 2019; **38**: 1276–1296.
13. Smeden MV, Moons KGM, Groot JAHD, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat Methods Med Res* 2019; **28**: 2455–2474.
14. Riley RD, Ensor J and Snell KIE. Calculating the sample size required for developing a clinical prediction model. *Br Med J* 2020; **368**: m441.
15. Riley RD, Van Calster B and Collins GS. A note on estimating the Cox-Snell R(2) from a reported C statistic (AUROC) to inform sample size calculations for developing a prediction model with a binary outcome. *Stat Med* 2021; **40**: 859–864.
16. De Jong VMT, Eijkemans MJC, Van Calster B, et al. Sample size considerations and predictive performance of multinomial logistic prediction models. *Stat Med* 2019; **38**: 1601–1619.

17. Van Smeden M, De Groot JAH, Moons KGM, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol* 2016; **16**: 1–12.
18. Nagelkerke NJD. A note on a general definition of the coefficient of determination. *Biometrika* 1991; **78**: 691–692.
19. Harrell FE. *Regression modeling strategies*. Cham, Switzerland: Springer, 2015.
20. Van Houwelingen J and Le Cessie S. Predictive value of statistical models. *Stat Med* 1990; **9**: 1303–1325.
21. Calster BV, Smeden MV, Cock BD, et al. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: simulation study. *Stat Methods Med Res* 2020; **29**: 3166–3178.
22. Riley RD, Snell KIE, Martin GP, et al. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *J Clin Epidemiol* 2021; **132**: 88–96.
23. Cox DR and Snell EJ. *Analysis of binary data*. 2 ed. London: Chapman and Hall/CRC, 1989.
24. Agresti A. *Categorical data analysis*. USA: Wiley Series, 2002.
25. Hoorde KV, Vergouwe Y, Timmerman D, et al. Assessing calibration of multinomial risk prediction models. *Stat Med* 2014; **33**: 2585–2596.
26. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**: 128–138.
27. Steyerberg EW. *Clinical prediction models - a practical approach to development, validation, and updating* (2nd ed). Cham: Springer, 2019.
28. Steyerberg EW and Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014; **35**: 1925–1931.
29. Begg CB and Gray R. Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika* 1984; **71**: 11–18.
30. Quesenberry CP and Hurst DC. Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics* 1964; **6**: 191–195.
31. Goodman LA. On simultaneous confidence intervals for multinomial proportions. *Technometrics* 1965; **7**: 247–254.
32. Van Calster B, Vergouwe Y, Looman CWN, et al. Assessing the discriminative ability of risk models for more than two outcome categories. *Eur J Epidemiol* 2012; **27**: 761–770.
33. Pate A. GitHub repository. Manchester predictive healthcare group. MRC-multi-outcome-project-8-multinomial-sample-size, [https://github.com/alexpate30/MRC-multi-outcome/tree/main/Project 8 Multinomial Sample Size](https://github.com/alexpate30/MRC-multi-outcome/tree/main/Project%208%20Multinomial%20Sample%20Size) (2021).
34. Timmerman D, Testa AC, Bourne T, et al. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group. *J Clin* 2005; **23**: 8794–8801. Epub ahead of print 2005. DOI: 10.1200/JCO.2005.01.7632.
35. Van Calster B, Van Belle V, Vergouwe Y, et al. Extending the c-statistic to nominal polytomous outcomes: the Polytomous Discrimination Index. *Stat Med* 2012; **31**: 2610–2626.
36. Austin PC and Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol* 2012; **12**. Epub ahead of print 2012. DOI: 10.1186/1471-2288-12-82.
37. Pencina MJ, Agostino RBD and Massaro JM. Understanding increments in model performance metrics. *Lifetime Data Anal* 2013; **19**: 202–218.
38. Su JQ and Liu JS. Linear combinations of multiple diagnostic markers linear combinations of multiple diagnostic markers. *J Am Stat Assoc* 1993; **88**: 1350–1355.
39. Zelen M and Severo NC. *Probability function*. Washington, DC: National Bureau of Standards Applied Mathematics, 1964.
40. Whitehead J. Sample size calculations for ordered categorical data. *Stat Med* 1993; **12**: 2257–2271.
41. Ensor J, Martin EC and Riley RD. pmsampsize: calculates the minimum sample size required for developing a multivariable prediction model. R package version 1.0.3, <https://cran.r-project.org/web/packages/pmsampsize/index.html> (2020).
42. Ensor J. PMSAMPsize: stata module to calculate the minimum sample size required for developing a multivariable prediction model.

## Appendix

### Nomenclature

CPM	clinical prediction model
$E_k$	number of individuals with outcome category $k$
$K$	number of outcome categories
LR	likelihood ratio test statistic
$m_{k,r}$	number of individuals with outcome category $k$ or $r$ that is required to target the shrinkage factor $S_{DL,k,r}$ to be above some pre-defined threshold
$n_{k,r}$	total number of individuals in the whole cohort that is required to target the shrinkage factor $S_{DL,k,r}$ to be above some pre-defined threshold ( $= m_{k,r} / p_{k,r}$ )



$p_k$	proportion of individuals that have outcome category $\in \{k\}$ ,
$p_{k,r}$	proportion of individuals that have outcome category $\in \{k, r\}$
$Q$	number of predictors parameters considered for inclusion prior to any variable selection
$R_{CS\_adj}^2$	optimism adjusted estimate of Cox-Snell's definition of explained variation for a binary logistic regression model
$R_{CS\_adj,k}^2$	$R_{CS\_adj}^2$ for distinct logistic regression model for sub-model $k$ , undefined reference category
$R_{CS\_adj,k,r}^2$	$R_{CS\_adj}^2$ for distinct logistic regression model for sub-model $k$ and reference category $r$
$R_{CS\_app}^2$	Apparent estimate of Cox-Snell's definition of explained variation for a binary logistic regression model
$R_{CS\_app,k}^2$	$R_{CS\_app}^2$ for distinct logistic regression model for sub-model $k$ , undefined reference category
$R_{CS\_app,k,r}^2$	$R_{CS\_app}^2$ for distinct logistic regression model for sub-model $k$ and reference category $r$
$R_{Nagelkerke}^2$	Nagelkerke's definition of explained variation ( $R^2$ ), $= R_{CS}^2 / \max(R_{CS}^2)$ , where $R_{CS}^2$ can be either $R_{CS\_adj}^2$ or $R_{CS\_app}^2$
$S$	global shrinkage factor for a binary logistic regression model
$S_{DL,k}$	shrinkage factor of distinct logistic regression model for sub-model $k$ , undefined reference category
$S_{DL,k,r}$	shrinkage factor of distinct logistic regression model for sub-model $k$ and reference category $r$
$S_{MN,k}$	sub-model specific shrinkage factor using the multinomial recalibration framework for sub-model $k$ , undefined reference category
$S_{MN,k,r}$	sub-model specific shrinkage factor using the multinomial recalibration framework for sub-model $k$ and reference category $r$
$S_{VH}$	heuristic shrinkage factor for a binary logistic regression model
$S_{VH\_MN}$	heuristic shrinkage factor for a multinomial logistic regression model
$\varphi_{k,r}$	outcome proportion in category $k$ relative to the reference category $r$ ( $= E_k / E_k + E_r$ )