

Supplemental Material:

**Compilation of Papers on Sample Size Requirements
for Discrete Choice Models**

by

Dominique Lord
Zachry Department of Civil and Environmental Engineering
Texas A&M University

Last Updated: April 12, 2026
(Added a small section on Machine Learning and AI models)

Note: This manuscript is an official supplemental material to the second edition of the textbook. It will become available on the Elsevier website associated with the textbook very shortly and it is now available on my own website associated with the textbook (link below). The citation for the textbook:

Lord, D., X. Qin, S.R. Geedipally, 2026. Highway Safety Analytics and Modeling, 2nd Ed. Elsevier Publishing Ltd., Amsterdam, NL.

https://dlord.engr.tamu.edu/wp-content/uploads/sites/234/2025/10/Sample_Size_Discrete_Choice_Models.pdf

Introduction

This manuscript presents a compilation of papers (and web links and a few textbooks) that discussed issues related to the required sample sizes (i.e., minimum sample size) for properly estimating logistic, multinomial logit, and mixed logit models as well as issues related to estimating the parameters (coefficients) of these models caused by small sample sizes. The papers come from a variety of fields, such as transportation or traffic engineering, industrial engineering, medical, biostatistics, marketing and econometrics among others. In many papers, the discussion about using a minimum sample was made in relation to preventing overfitting parameters (i.e., the sample size is too small in relation to the number of parameters and levels included in the model¹), usually with the goal of adequately estimating or predicting the population-based risk or outcome (Steyerberg, 2009). When overfitting is detected or identified, most authors recommend using penalty functions, such as the Least Absolute Shrinkage and Selection Operator (LASSO), Ridge Regression, and Firth's correction, although, in recent years, some researchers have recommended using an adequate sample size over penalizing the models. These penalty functions are used to shrink the estimated parameters to reduce or eliminate overfitting² (see de Jong et al., 2019, for example). The compilation shows that estimating the required sample size is complex, but the consensus indicates that considering the sample size for estimating the parameters of these models is very important. The compilation may, on occasion, be updated as more relevant references from the literature are identified and added to this manuscript. At the date shown on the title page, 65 papers or textbooks were included in this manuscript. These papers only represent the tip of the iceberg for all the work that focused on estimating or using adequate sample sizes or identifying issues with a small sample size for discrete choice models.

This manuscript first covers the important points or conclusions identified from the compilation related to the sample size requirements and discrete choice models. Then, the second section alphabetically lists the papers and textbooks collected for this work. The last section provides excerpts from these papers and textbooks as well as comments about the content of the paper when needed.

Summary

Table 1 summarizes different values and methods that have been proposed for determining sample size requirements for binary logit or logistic, multinomial logit and mixed logit models.

Table 1. Summary of Proposed Values for the Required Minimum Sample Size

Modeling Type	Sample Size
Binary Logit/Logistic	<p data-bbox="522 1650 797 1682"><u>Traditional approach:</u></p> <p data-bbox="522 1728 1373 1795">10 Events per Variable/Parameter (EPV) is the absolute minimum. A value equal to 20 is preferable. These values apply to the smallest</p>

¹ https://statisticsbyjim.com/regression/overfitting-regression-models/#Ogoogle_vignette

² <https://medium.com/@jainvidip/ridge-and-lasso-the-dynamic-duo-of-regularization-a5d22d903bab>

	<p>category or less frequent outcome of the two. If the binary logit model contains 15 variables for instance, the minimum sample size should be 150 observations, with 300 observations being preferable, for the smallest category. These values are proposed for the general estimation of the parameters (i.e., stability or biases related to the parameter estimates) in the logistic model (without specifically looking at overfitting [i.e., quantifying the amount of overfitting]).</p> <p>Note: For many researchers, an EPV fewer than 10 is considered a small sample size.</p> <p><u>Other approaches or recommendations:</u></p> <p>Hosmer and Lemeshow (2000) have devoted almost an entire chapter about the required minimum sample sizes in their textbook on logistic regression. They provided different methods or equations that are dependent on the number of parameters in the model, significance level (say at the 5% level), the study power, odds ratio and the proportion of the outcome variable. It is unclear if these methods are still used for estimating the required sample size. In their example applications, the sample sizes varied from 500 to 1,000 observations (total sample).</p> <p>Using simulation, Ogundimu et al. (2016) suggested that the EPV should be larger than 20, especially if the smallest category has very small outcome or low proportion. Looking at the total sample size should also be considered.</p> <p>Also using simulation, van Smeden et al. (2016) indicated that biased parameter estimates can be observed for EPVs as large as 40 to 50 (important enough), but like Ogundimu et al. (2016), other factors, such as the data split/proportion, that influence the required sample size should be examined. The authors of both papers showed that small sample sizes led to biased parameter estimates.</p> <p>Bujang et al. (2023) recommended using a minimum of 500 observations irrespective of the split ratio between the two categories in the binary logit model. However, they nonetheless still strongly advise using an NPV equal to 50.</p>
<p>Multinomial Logit (MNL)</p>	<p><u>Traditional approach:</u></p> <p>Like the binary logit model, 10 Events per Variable (EPV) is the absolute minimum. This also applies to the smallest category or less frequent</p>

	<p>outcome among all the categories. A value equal to 20 is preferable. If the multinomial logit model contains 15 variables, the minimum number of observations should be 150, with 300 observations being preferable, for the smallest category (say fatalities in a crash-severity model).</p> <p><u>Other approaches or recommendations:</u></p> <p>Some studies have recommended a minimum of 1,000 (Koppelman and Chu, 1983), 2,000 (Ye and Lord, 2014) and from 3,000 to 4,000 observations (Zeng et al., 2018) (total sample). Like van Smeden et al. (2016) and Ogundimu et al. (2016) among others, the authors of the three papers examined the stability of the parameters as a function of sample sizes. All these authors showed that small sample sizes led to biased or unreliable parameters.</p> <p>de Jong et al. (2019) suggested that the EPV should be equal to 50. However, the EPV can be reduced if a ridge regression or the LASSO penalized technique is used to decrease, if not eliminate, overfitting particularly for an EPV below 10. These authors specifically quantified overfitting. In the example above, an EPV of 50 would mean that 750 fatalities (observations) would be needed in the smallest group.</p> <p>van Calster et al. (2020), on the other hand, have examined several penalizing functions (Ridge Regression, LASSO, Adaptive LASSO, Firth's correction, etc.) and noted that these functions may not always reduce overfitting because they are estimates themselves. Riley et al. (2021b) further confirmed this and noted that penalizations and shrinking methods still do not eliminate overfitting for multinomial logit models when the sample size is small or very small (an EPV below 10). The authors of both papers still recommend continuing using an EPV=50 or not using penalty functions at all (that is using an adequate sample size, which is the best option).</p> <p>One study (Hamzah et al., 2016) indicated that positively and negatively skewed variables as well as categorical variables require larger sample sizes. Only including one such variable required at least 300 observations for positively skewed and categorical variables. A larger sample was needed for negatively skewed parameters. Along the same line, de Jong et al. (2019) noted that categorical variables were the ones that greatly affected the overfitting of the data (as opposed to continuous variables).</p>
--	--

	<p>The following five papers are the critical manuscripts recommended for estimating or validating the required sample size: de Jong et al. (2019), Riley et al. (2019b, 2021a), Pate et al. (2023), and Gehringer et al. (2024b). All these papers combined have been cited more than 1,500 times in SCOPUS. The proposed methods can also be used for logistic regression as well as for mixed logit models (the methods documented in de Jong et al. [2019] and Pate et al. (2023), for example, can be utilized to assess overfitting).</p> <p>Example 4.2 in the textbook describes an application of the methods to identify overfitting in both multinomial logit and mixed logit models.</p> <p>Gehringer et al. (2024a) explained how to determine the maximum number of variables that can be used in a logistic regression/MNL model when the number of available observations is fixed or given.</p>
<p>Mixed Logit</p>	<p><u>Traditional approach:</u></p> <p>There is currently no traditional approach for estimating the required sample size, although the traditional approach for the MNL model may be applicable here at the bare minimum.</p> <p><u>Other approaches or recommendations:</u></p> <p>Ye and Lord (2014) recommended a minimum of 5,000 observations. Smaller sample sizes led to biased estimates. The recommendation was based on a model with 7 parameters and 5 levels. As described above, the analysis did not quantify overfitting.</p> <p>Although Wang et al. (2021) did not specifically examine sample size requirements in their comparison of deep neural networks with multinomial logit and mixed logit models, their simulation work showed that multinomial logit and mixed logit models with 20 and 50 variables did not become stable in terms of prediction error (i.e., overfitting) until 4,000 to 5,000 observations, with the multinomial logit converging faster. Both models were unreliable with observations below 3,000.</p> <p>Assele et al. (2023) noted in their study on the required minimum sample size for MNL models: <i>“Last but not least, it is very important to note that the study was restricted to the MNL model whereas the mixed logit model, latent class model, nested logit model, or other extensions of the MNL model will generally require a larger sample.”</i></p>
<p>Machine Learning/</p>	

Artificial Intelligence	<p><u>Traditional approaches:</u></p> <p>There are no traditional approaches for determining the required sample sizes for these kinds of models (e.g., xGBoost, Neural Networks, Random Forest, Support Vector Machine). See Chapter 12 for an extensive list.</p> <p><u>Other approaches or recommendations:</u></p> <p>The research on the minimum or required sample sizes for these models is on-going given how recently they were introduced, their complexity (Melvin, 2021; Thankappan, 2024) and how many models exist. In general, like the other models described above, the common theme consists of minimizing or eliminating overfitting when the models are developed. This category of models may be the subject of another supplemental material in the near future. Some recent highlights include:</p> <p><i>“Sophisticated models overfitted in small datasets but maximized holdout test results in larger datasets.”</i> (Zantvoork et al., 2024)</p> <p><i>“To reinforce this recommendation, we outline seven reasons why inadequate sample size negatively affects model training, evaluation, and performance.”</i> (for AI models) (Riley et al., 2025)</p> <p><i>“We proposed the minimum sample size for the neural network model fitting with two criteria: the performance of 95% of the models is close to the theoretical maximum, and 80% of the models can outperform the linear model.”</i> (Cheng et al., 2025)</p> <p><i>“Although model performance metrics were similar across models, substantial differences in effective sample sizes and risk predictions were observed among patients in the clinical dataset.”</i> (Thomassen et al., 2025)</p> <p><i>“Only 30% of studies met the minimum sample size required for ML models, indicating potential overfitting and limited generalizability.”</i> (Zamagni et al., 2026)</p> <p><i>“Comparison of results to sample sizes obtained from differential analysis power analysis methods showed that ML methods generally required larger sample sizes.”</i> (Silvey et al., 2026)</p>
--------------------------------	--

Here are the important highlights:

- All the papers/manuscripts or web links clearly indicate that a minimum required sample size is needed for properly estimating discrete choice models and small sample sizes lead to biased or unreliable parameter estimates. In fact, the issue related to the small sample bias is true for any statistical distribution or model³⁴.
- Martin et al. (2025) reported the following about multinomial logit models: “*It is crucial to ensure that the sample size of the data used to develop or validate a clinical prediction model⁵ is large enough. If the data are inadequate, developed models can be unstable and estimates of predictive performance imprecise. This can lead to models that are unfit or even harmful for clinical practice.*”
- Zaloumis et al. (2025) indicated that “It is common for no sample size calculations or assessments to be performed a priori in prediction modelling studies, and as Riley et al. [1, 2] highlight, in clinical research more broadly, many prediction models are developed using datasets (commonly referred to as the training dataset) that are too small for the number of participants and outcome events.”

³ Many researchers have devoted a major part of their career examining issues related to small sample size and the development of statistical models or estimating the required sample sizes. See Dr. Lawrence Joseph (<https://joseph.research.mcgill.ca/Research-Interests.html>) (he created nine programs in R, WinBUGS and Perl to calculate the required sample sizes for different models, including the logistic regression: <https://joseph.research.mcgill.ca/software/Bayesian-Sample-Size.html>) and several of the researchers listed in this manuscript.

⁴ Interestingly, Dr. Milica Miočević has recently co-edited a textbook on this topic (<https://www.routledge.com/Small-Sample-Size-Solutions-A-Guide-for-Applied-Researchers-and-Practitioners/vandeSchoot-Miocevic/p/book/9780367222222>) (“Researchers often have difficulties collecting enough data to test their hypotheses, either because target groups are small or hard to access, or because data collection entails prohibitive costs. Such obstacles may result in data sets that are too small for the complexity of the statistical model needed to answer the research question. This unique book provides guidelines and tools for implementing solutions to issues that arise in small sample research.”).

⁵ The terminology “Clinical Prediction Model” is like the terminology used in highway safety such as “Crash Prediction Model,” which is synonymous to “Safety Performance Function” or “Crash-Frequency Model” for count data models. Irrespective of the terminology used, “Clinical Prediction Models” (often used to estimate or predict disease-related risk or outcome as well as unintentional injuries—which include those caused by motor vehicle crashes—, as described in the introduction) are estimated in the same manner as any discrete choice models. As stated by Gehringer et al. (2024b) in their paper describing how to develop and validate such models, they are developed with the goal of properly estimating the coefficients of the model, as also noted by van Smeden et al. (2016) and many others: “*To develop a multinomial model, a MLR (multinomial logit regression) is fitted using maximum likelihood estimation (Box 1). The model could also be fit using penalised regression approaches (23), such as LASSO, ridge methods, or Firth’s correction (58,59). These may help limit overfitting when coupled with appropriate sample sizes (60). However, the shrinkage parameter itself requires estimation (59,61) so ideally the sample size should be sufficient to not need shrinkage. Continuous predictors should not be categorized during model fitting, to avoid loss of information (62,63).*” (Note: ‘appropriate’ above refers to medium sample size and the authors recommend not developing MLRs with a small sample size because many penalizing functions may not work.)

- Depending on the characteristics of the data, model attributes (e.g., the number of levels and their proportions or ratios) and study objectives, the minimum sample size can be extremely large for the traditional multinomial logit model, as listed in the literature below. For example,
 - Pate et al. (2023) wrote the following: *“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.”*
 - Hippisley-Cox and Coupland (2025) noted the following for selecting the minimum required sample size (MNL models with 60 to 100 variables): *“Across these criteria the sample size of the available data met the required minimum sample sizes of 3,382,393 in women and 3,227,602 in men for the specified criteria, allowing for a reduced shrinkage target for the rare cancer pairs where there may be some overfitting.”* (Note: in the supplemental material, the authors reported required sample sizes varying from 11,000 observations to close to 1 million observations for other discrete choice models, as shown in the third section below.)
 - Elayan et al. (2025) wrote the following in their study involving case mixed-shift for model development: *“Furthermore, the minimum sample size criteria (sic) was calculated using the method of Riley et al. [22], where we assumed an event prevalence of 0.0021, a target shrinkage factor of 0.9, number of predictors of 7, and c-statistic of 0.75 [40], this gave a required sample size of 33,123. This minimum sample size criteria (sic) was met during the models’ development.”*
- Assele et al. (2023) noted the following in the conclusions of the paper: *“Last but not least, it is very important to note that the study was restricted to the MNL model whereas the mixed logit model, latent class model, nested logit model, or other extensions of the MNL model will generally require a larger sample.”*
- Hamzah et al. (2016) indicated that positively and negatively skewed as well as categorical variables in multinomial logit models require larger sample sizes (negatively skewed required larger sample size than positively skewed). That study examined one variable at a time, and the authors found a minimum number of observations equal to 300 when only one variable was evaluated.
- Like the previous point, de Jong et al. (2019) noted that categorical variables had a greater effect on the overfitting of multinomial logit models than continuous variables when the sample size was below the recommended value.
- Wang et al. (2024) analyzed and compared more than 100 machine learning and discrete choice models and noted the following: *“Firstly, many ML models,*

particularly the ensemble methods and deep learning, statistically outperform the DCM family and its individual variants (i.e., multinomial, nested, and mixed logit), thus corroborating with the previous research. However, this study also highlights the crucial role of the contextual factors (i.e., data sources, inputs and choice categories), which can explain models' predictive performance more effectively than the differences in model types alone. Model performance varies significantly with data sources, improving with larger sample sizes and lower dimensional alternative sets."

- None of the studies, from biostatistics, statistics to engineering, have used the likelihood ratio test (LRT) to make decisions about the minimum or required sample size requirements. More advanced statistical tests were utilized for such purposes. They included using the calibrating slopes, Brier score, Nagelkerke R^2 , $R^2_{CS, adj}$ (Cox-Snell R-squared), and discriminant analysis among others. These tests can be used to identify or quantify overfitting⁶ (Hensher et al., 2013⁷; Rigg and Hankins, 2015; Zhao et al., 2019⁸; Zabor et al., 2022⁹; Nibbering, 2024¹⁰) and eventually trying to correct the coefficients (i.e., shrinking the coefficients) using the LASSO technique and Ridge Regression (since the LRT has difficulties identifying overfitted

⁶ "Overfit regression models have too many terms for the number of observations. When this occurs, the regression coefficients represent the noise rather than the genuine relationships in the population." (<https://statisticsbyjim.com/regression/overfitting-regression-models/>) See also <https://blog.minitab.com/en/blog/adventures-in-statistics-2/the-danger-of-overfitting-regression-models>, https://tung-dn.github.io/prog_class3.html and <https://medium.com/@sahin.samia/mregularization-in-regression-tackling-overfitting-for-enhanced-model-performance-553f2cef02df> (*The overfitted model essentially learns the noise in the data rather than the actual underlying trend.*).

⁷ Dr. William H. Green has provided a warning about overfitting mixed logit models with latent classes "Signature features of a model that has been overfit include exceedingly small estimates of the class probabilities, wild values of the structural parameters, and huge estimated standard errors."

⁸ "Finally, it is somewhat surprising that the mixed logit model, a model that accounts for individual heterogeneity and has significantly better model fit (adjusted McFadden's pseudo R^2 is 0.536) than the MNL model (adjusted McFadden's pseudo R^2 is 0.365), underperformed the MNL model in terms of the out-of-sample predictive power. This finding is nonetheless consistent with the findings of Cherchi and Cirillo (2010). It suggests that the mixed logit model may have overfitted the data with the introduction of random parameters, and such overfitting resulted in greater out-of-sample prediction error."

⁹ "A logistic regression model that is overparameterized (i.e., too many variables for too few events) can result in odds ratios that are implausibly large and confidence intervals that are wide and uninterpretable. These types of "overfitted" models should be avoided."

¹⁰ "With a large choice set, the number of parameters in the $J \times K$ matrix β is large. Large numbers of parameters amplify overfitting concerns and make it difficult to extract useful insights. For the data to be informative on the parameters without additional restrictions, the number of outcome categories and explanatory variables need to be relatively small."

models^{11 12 13 14}; in fact, the LRT becomes unreliable when the number of parameters is too large with respect to the sample size¹⁵).

- Almost all the papers described below have used simulation protocols¹⁶ to evaluate the negative effects of small sample sizes and their influence on the performance of discrete choice models. Using simulation in traffic safety research to evaluate new distributions and models or examining methodological issues such as small sample sizes is no different than the simulations documented in the papers identified in this manuscript. The researchers of the papers covered in this manuscript understand how to conduct simulation. There are no issues with “mis-specifications” both in terms of simulated data and with developing discrete choice models with the purpose of properly estimating coefficients (see, e.g., Riley¹⁷ et al. [2021b] who used simulation to specifically identify issues with small sample sizes and the performance of multinomial logit models as a function of penalizing functions [*“Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small”*]). This topic is also covered in Chapter 6 of the textbook.

The effects of small samples on the estimated coefficients are illustrated in Figure 1. This figure shows the characteristics of a hypothetical coefficient based on a very large sample and two subsets of the original sample. One coefficient is misestimated because the model is considered overfitted whereas the other would be removed from the model although it should not be. As discussed by Babuak (2004), overfitted models can produce unreliable and/or biased coefficients, coefficients that are too large (hence, using the

¹¹ “Evaluate the implications of using likelihood ratio tests in large sample sizes versus small sample sizes within continuous distributions. Using likelihood ratio tests in large sample sizes tends to yield significant results even for minor differences in model fit, which can lead to overfitting or misleading conclusions. Conversely, in small samples, these tests may lack power to detect meaningful differences, resulting in potential Type II errors. This contrast highlights the need for researchers to consider sample size when interpreting likelihood ratio test outcomes and selecting appropriate models for continuous distributions.” Fiveable. “Likelihood Ratio Tests – Intro to Probability.” Edited by Becky Bahr, Fiveable, 2024, <https://library.fiveable.me/key-terms/introduction-probability/likelihood-ratio-tests>. The logistic regression is a continuous distribution. On the other hand, the multinomial logit model is considered a discrete distribution, but the LRT test is still affected by the sample size; see He et al. (2021) who looked at multinomial logit models; Yuan et al. (2019) for structural equation modeling. Other references on this topic include Royle and Dozario (2009), Hughes (2017) and Rold and Sidaty-Regad (2021).

¹² <https://statisticsbyjim.com/regression/overfitting-regression-models/>

¹³ <https://stats.stackexchange.com/questions/213473/does-likelihood-ratio-test-control-for-overfitting>

¹⁴ <https://www.geeksforgeeks.org/r-language/likelihood-ratio-test/>

¹⁵ “High-Dimensional Data: In scenarios with many variables, the chi-squared approximation can become inaccurate, leading to misleading conclusions,” as discussed in Bai et al. (2009), He et al. (2020, 2021) and the references herein.

¹⁶ In many cases, the authors used real data to demonstrate or support the simulation results.

¹⁷ His work has been cited more than 58,400 times with an h-index of 111, according to Google Scholar.

LASSO for example) or have a high variance (in repeated subsamples, the value of the coefficient significantly changes from one sample to the next) (this author included logistic regression in the discussion) (This is also discussed by Lever et al., 2016¹⁸). This author clearly suggests that increasing the sample size will help overcome these issues.

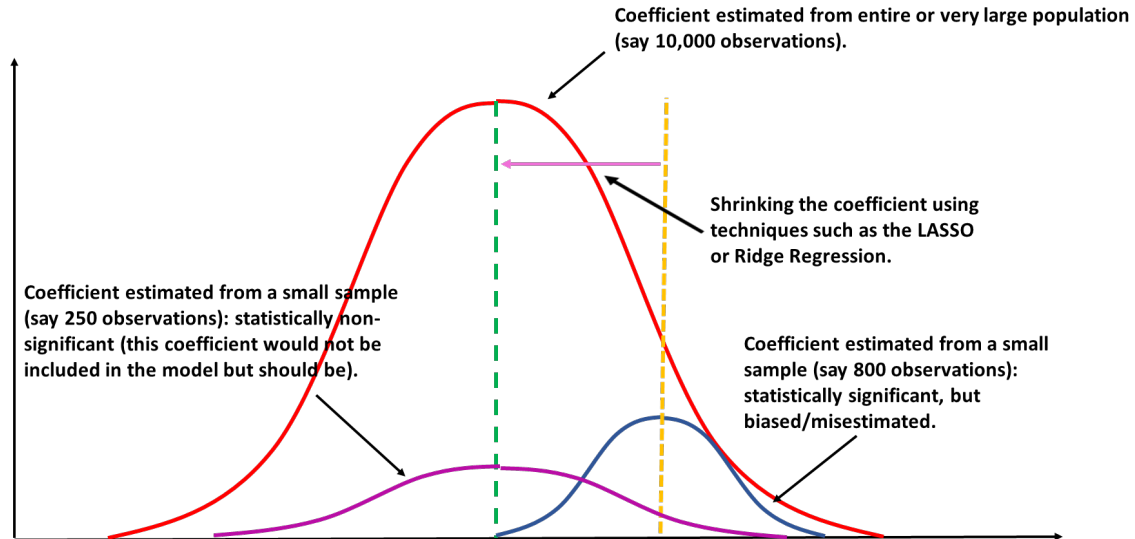


Figure 1. Characteristics of a coefficient based on different sample sizes.

In his excellent book on developing clinical prediction models, Steyerburg (2009) devotes an entire chapter (and several other sections) on issues related to overfitting regression models (the focus is on discrete choice models). In his words, “Overfitting is a major problem in regression modelling. It arises from two main issues: model uncertainty and parameter uncertainty.” (p. 88) He describes in great details how overfitting is a critical issue, not only for prediction purposes (population risk), but also for the interpretation of the coefficients or parameters (i.e., the overestimation effects of the parameters, as he puts it), since, as discussed above, the coefficients are capturing noises rather than the true relationship between the dependent and independent variables. He illustrates this using the two figures shown below. Figure 2 shows the biased estimate of an overfitted parameter on the right, while Figure 3 explains how the Mean Squared Error (used as an approach to estimate overfitting) changes as a function of the number of parameters included in the model. In this figure, internal performance refers to using the sample population from which the parameters were estimated. On the other hand, external performance shows how the model performs for predicting estimates.

¹⁸ Used the well-known quote by John von Neumann: “With four parameters I can fit an elephant and with five I can make him wiggle his trunk.”

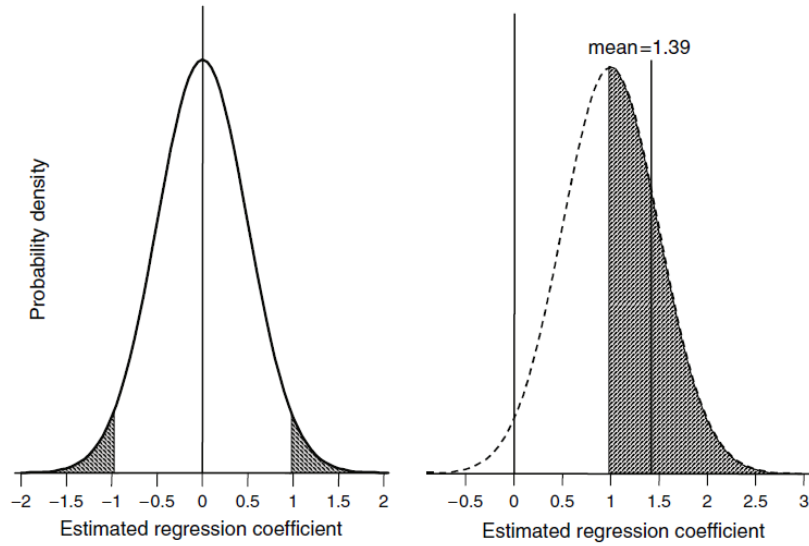


Fig. 5.5 Illustration of testimation bias. In case of a noise variable, the average of estimated regression coefficients is zero, and 2.5% of the coefficients is below -0.98 ($1.96 \times SE$ of 0.5), and 2.5% of the coefficients is larger than $+0.98$ ($1.96 \times SE$ of 0.5). In case of a true coefficient of 1, the estimated coefficients are statistically significant in 52%. For these cases, the average of estimated coefficients is 1.39 instead of 1

Figure 2. Illustration of an overfitted parameter (Steyerberg, 2009).

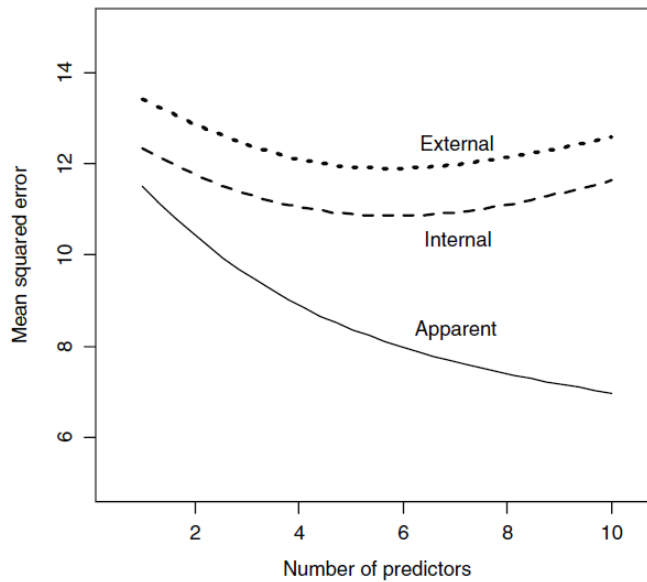


Fig. 5.6 Mean squared error of predictions from models with increasing complexity (1,000 simulated samples with $n = 50$). Apparent performance improves with more predictors, but internal and external performances worsen with more than five predictors

Figure 3. Illustration of the Mean Squared Error as a function of the number of parameters in the model (Steyerberg, 2009).

Critical Papers for Estimating or Validating Required Sample Sizes

- de Jong, V.M.T., M.J.C. Eijkemans, B. van Calster, D. Timmerman, K.G.M. Moons, E.W. Steyerberg, M. van Smeden., 2019. **Sample size considerations and predictive performance of multinomial logistic prediction models.** *Statistics in Medicine*, Vol. 38, No. 9, pp. 1601-1619. <https://doi.org/10.1002/sim.8063>
- Pate, A., R.D. Riley, G.S. Collins, M. van Smeden, B. Van Calster, J. Ensor, G.P. Martin, 2023. **Minimum sample size for developing a multivariable prediction model using multinomial logistic regression.** *Statistical Methods in Medical Research*. Vol. 32, No. 3, 555–571. <https://doi.org/10.1177/09622802231151220>

You can find the R codes for determining the minimum sample size here: [Project 8 Multinomial Sample Size](#)

- Riley, R.D., K.I. Snell, J. Ensor, D.L. Burke, F.E. Harrell Jr., K.G.M. Moons, G.S. Collins, 2019b. **Minimum sample size for developing a multivariable prediction model: PART II – binary and time-to-event outcomes.** *Statistics in Medicine*. 38, 1276–1296. (minor correction: Correction to: Minimum sample size for developing a multivariable prediction model: Part II-binary and time-to-event outcomes by Riley RD, Snell KI, Ensor J, et al. <https://doi.org/10.1002/sim.8409>)
- Riley, R.D., T.P.A. Debray, G.S. Collins, L. Archer, J. Ensor, M. van Smeden, K.I.E. Snell, 2021a. **Minimum sample size for external validation of a clinical prediction model with a binary outcome.** *Statistics in Medicine*, Vol. 40, No. 19., pp. 4230-4251. <https://doi.org/10.1002/sim.9025> (This paper focuses on describing a step-by-step procedure for applying an existing logistic model to a new or independent dataset to externally validate the model. In this case, the required sample size is used to ensure the outcome of the application to a new dataset is valid and, if invalid, it is not caused by a sample size that is too small. Note that the recently published paper by Sadatsafavi et al. (2026), listed below, describes similar steps, but for a logistic model estimated using the Bayesian method.)
- Gehringer, C.K., G.P. Martin, B. Van Calster, K.L. Hyrich, S.M.M. Verstappen, J.C. Sergeant, 2024b. **How to develop, validate, and update clinical prediction models using multinomial logistic regression.** *Journal of Clinical Epidemiology*, Vol. 174, 111481. <https://doi.org/10.1016/j.jclinepi.2024.111481>

On the publisher's website, you can access the step-by-step instructions for estimating the required sample size for a multinomial logit model. (<https://ars.els-cdn.com/content/image/1-s2.0-S0895435624002373-mmc1.docx>)

This paper describes a reversed approach (i.e., estimating the maximum number of parameters that can be included in a model when the sample size is fixed or given).

- Gehringer, C.K., G.P. Martin, K.L. Hyrich, S.M.M. Verstappen, J. Sextone, E.K. Kristianslund, S.A. Provane, T.K. Kvien, J.C. Sergeant, 2024a. **Developing and externally validating multinomial prediction models for methotrexate treatment outcomes in patients with rheumatoid arthritis: results from an international collaboration.** Journal of Clinical Epidemiology, Vol. 166, 111239. <https://doi.org/10.1016/j.jclinepi.2023.111239>

Papers Listed Alphabetically

1. Abonazel, M.R., I. Dawoud, M. Naji Al-Ghamdi, R.A. Farghali, 2025. Developing the generalized Dawoud-Kibria estimator for the multinomial logistic model: Simulation study and application. Scientific African, Vol. 29, e02803. <https://doi.org/10.1016/j.sciaf.2025.e02803>
2. Archer, L., K.I.E. Snell, J. Ensor, M.T. Hudda, G.S. Collins, R.D. Riley, 2021. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. Statistics in Medicine, Vol. 40, No. 1, pp. 133-146. DOI: 10.1002/sim.8766 (This paper is not directly related to discrete choice models but uses the same methods for determining the required sample sizes.)
3. Assele, S.Y., M. Meulders, M. Vandebroek, 2023. Sample size selection for discrete choice experiments using design features. Journal of Choice Modelling, 49, 100436. DOI: 10.1016/j.jocm.2023.100436.
4. Baeza-Delgado, C. L.C. Alberich, J.M. Carot-Sierra, D. Veiga-Canuto, B. M. de las Heras, B. Raza, L. Martí-Bonmatí, 2022. A practical solution to estimate the sample size required for clinical prediction models generated from observational research on data. European Radiology Experimental, Vol. 6, No. 22. (<https://doi.org/10.1186/s41747-022-00276-y>)
5. Biscadi, S., J. Orprecio, M.B. Fenton, A. Tsoar, J.M. Ratcliff, 2004. Data, sample sizes and statistics affect the recognition of species of bats by their echolocation calls. Acta Chiropterologica, 6(2): 347–363. <https://doi.org/10.3161/001.006.0212>

6. Bujang MA, Sa'at N, Tg Abu Bakar Sidik TMI, Lim CJ., 2023. Sample size guidelines for logistic regression from observational studies with large population: emphasis on the accuracy between statistics and parameters based on real life clinical data. *Malays J Med Sci.*; 25(4):122–130. <https://doi.org/10.21315/mjms2018.25.4.12>
7. Cao, Q., Z. Jiang, Z. Wang, L. Wee, A. Dekker, Z. Zhang, J. Zhu, 2025. Minimum sample size calculation for radiomics-based binary outcome prediction models: Theoretical framework and practical example. *Radiotherapy and Oncology*, Vol. 212, 111134. <https://doi.org/10.1016/j.radonc.2025.111134>
8. Collins, G.S., P. Dhiman, J. Ma, M.M. Schlusset, L. Archer, B. Van Calster, F.E. Harrell Jr, G.P. Martin, K.G.M Moons, M. van Smeden, M. Sperrin, G.S. Bullock, R.D. Riley, 2024. Evaluation of clinical prediction models (part 1): from development to external validation. *British Medical Journal*, Vol. 384, e074821. <http://dx.doi.org/10.1136/>
9. de Bekker-Grob, E.W., B. Donkers, M.F. Jonker, E.A. Stolk, 2015. Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide. *Patient*, Vol. 8, 373–384.
10. de Jong, V.M.T., M.J.C. Eijkemans, B. van Calster, D. Timmerman, K.G.M. Moons, E.W. Steyerberg, and M. van Smeden., 2019. Sample size considerations and predictive performance of multinomial logistic prediction models. *Statistics in Medicine*, Vol. 38, No. 9, 1601-1619. DOI: 10.1002/sim.8063
11. Dekker, T., P. Bansal, J. Huo, 2025. Revisiting McFadden's correction factor for sampling of alternatives in multinomial logit and mixed multinomial logit models. *Transportation Research Part B*, 103-129. <https://doi.org/10.1016/j.trb.2024.103129>
12. Dhiman, P., J. Ma, C. Qi, G. Bullock, J.C. Sergeant, R.D. Riley, G.S. Collins, 2023. Sample size requirements are not being considered in studies developing prediction models for binary outcomes: a systematic review. *BMC Medical Research Methodology*, Vol. 23, No. 188. <https://doi.org/10.1186/s12874-023-02008-1>.
13. Diomatari, C., G.P. Martin, D. A. Jenkins, M. Jani, 2025. Clinical prediction models for medication adverse events in patients with rheumatic and musculoskeletal conditions: A systematic literature review. *Seminars in Arthritis and Rheumatism*, Vol. 73, 152728.
14. Edlinger, M., M. van Smeden, H.F. Alber, M. Wanitschek, B. Van Calster, 2022. Risk prediction models for discrete ordinal outcomes: Calibration and the impact of the

proportional odds assumption. *Statistics in Medicine*, Vol. 41, 1334–1360. DOI: 10.1002/sim.9281.

15. Elayan, H., M. Sperrin, G.P. Martin, N. Peek, F. Braunschweig, J. Faxén, J. Alfredsson. D.A. Jenkins, 2025. Correcting for case-mix shift when developing clinical prediction models. *BMC Medical Research Methodology*, Vol. 25, 186.
<https://doi.org/10.1186/s12874-025-02621-2>
16. Falke, A., H. Hruschka, 2017. Setting prices in mixed logit model designs. *Marketing letters*, Vol. 28, 139–154: DOI 10.1007/s11002-015-9396-4
17. Farghali, R.A., M. Qasim, B.M. Golam Kibria, M.R. Abonazel. 2023. Generalized two parameter estimators in the multinomial logit regression model: methods, simulation and application. *Communications in Statistics - Simulation and Computation*, 52:7, 3327-3342, DOI: 10.1080/03610918.2021.1934023
18. Fuetterer, C., M. Nalenz, T. Augustin, R.M. Pfeiffer, 2025. A powerful penalized multinomial logistic regression approach. *Computational Statistics*.
<https://link.springer.com/article/10.1007/s00180-025-01635-0> (open source)
19. Gehringer, C.K., G.P. Martin, K.L. Hyrich, S.M.M. Verstappen, J. Sextone, E.K. Kristianslund, S.A. Provane, T.K. Kvien, J.C. Sergeant, 2024a. Developing and externally validating multinomial prediction models for methotrexate treatment outcomes in patients with rheumatoid arthritis: results from an international collaboration. *Journal of Clinical Epidemiology* 166 (2024) 111239.
<https://doi.org/10.1016/j.jclinepi.2023.111239> (This paper described a reversed approach. The researchers only had a fixed sample size of 1,632 observations and could therefore only fit a model with 8 parameters.)
20. Gehringer, C.K., G.P. Martin, B. Van Calster, K.L. Hyrich, S.M.M. Verstappen, J.C. Sergeant, 2024b. How to develop, validate, and update clinical prediction models using multinomial logistic regression. *Journal of Clinical Epidemiology* 174 (2024) 111481. <https://doi.org/10.1016/j.jclinepi.2024.111481>
21. Hamzah, Y.B.W., X.-J. Xie, 2016. Effects of different type of covariates and sample size estimation for multinomial logistic regression model. *Jurnal Teknologi (Sciences & Engineering)*, Vol. 78 12–3, 155–161.
22. Heckmann, T., K. Gegg, A. Gegg, M. Becht, 2014. Sample size matters: investigating the effect of sample size on a logistic regression susceptibility model for debris flows. *Natural Hazards and Earth System Sciences*, Vol. 14, 259–278.
doi:10.5194/nhess-14-259-2014

23. Hippisley-Cox, J., C.A. Coupland, 2025. Development and external validation of prediction algorithms to improve early diagnosis of cancer. *Nature Communications*, Vol. 16, 3660. <https://doi.org/10.1038/s41467-025-57990-5>
24. Hossain, M.B., M. Sadatsafavi, J.C. Johnston, H. Wong, V.J. Cook, M.E. Karim, 2025. LASSO-Based Survival Prediction Modeling with Multiply Imputed Data: A Case Study in Tuberculosis Mortality Prediction. *The American Statistician*. Vol. 00, No. 0, 1–12: Statistical Practice. <https://doi.org/10.1080/00031305.2025.252654>
25. Jahan, M.D., T. Bhowmik, L. Hoover, N. Eluru, 2025. Comparing the Performance of Different Missing Data Imputation Approaches in Discrete Outcome Modeling. *Transportation Research Record*, Vol. 2679(2), 879–903.
26. Kim, S., E. Heath, L. Heilbrun, 2017. Sample size determination for logistic regression on a logit-normal distribution. *Statistical Methods in Medical Research*, Vol. 26(3), 1237–1247. <https://doi.org/10.1177/0962280215572407>
27. Koppelman, F.S., C. Chu, 1983. Effect of sample size on disaggregate choice model estimation and prediction. *Transportation Research Record No. 944*, 60–69. <https://onlinepubs.trb.org/Onlinepubs/trr/1983/944/944-009.pdf>
28. Legha, A., J. Ensor, R. Whittle, L. Archer, B. Van Calster, E. Christodoulou, K.I.E. Snell, M. Sadatsafavi, G.S. Collins, R.D. Riley, 2026. Sequential sample size calculations and learning curves safeguard the robust development of a clinical prediction model for individuals. *Journal of Clinical Epidemiology*, Vol. 191, 112117.
29. Månsson, K., G. Shukur, B.M. Golam Kibria, 2018. Performance of some ridge regression estimators for the multinomial logit model, *Communications in Statistics - Theory and Methods*, Vol. 47, No. 12, 2795–2804, DOI: 10.1080/03610926.2013.784996
30. Martin, G.P., M. Sperrin, K.I.E. Snell, I. Buchan, R.D. Riley, 2021. Clinical prediction models to predict the risk of multiple binary outcomes: a comparison of approaches. *Statistics in Medicine*. Vol. 40, 498–517. DOI: 10.1002/sim.8787
31. Martin, G.P., R.D. Riley, J. Ensor, S.W. Grant, 2025. Statistical primer: sample size considerations for developing and validating clinical prediction models. *European Journal of Cardio-Thoracic Surgery*, Vol. 67, No. 5, doi:10.1093/ejcts/ezaf142.
32. Nemes, S., J.M. Jonasson, A. Genell, G. Steineck, 2009. Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology*, Vol. 9, No. 56. doi:10.1186/1471-2288-9-56

33. Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, A.R. Feinstein, 1996. A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. *Clinical Epidemiology*, Vol. 49, No. 12, pp. 1373–1379.
<https://doi.org/10.1016/j.amepre.2003.12.002>
34. Ogundimu, E.O., D.G. Altman, G.S., Collins, 2016. Adequate sample size for developing prediction models is not simply related to events per variable. *Journal of Clinical Epidemiology*, Vol. 76, 175-82.
35. Orme, B. 2019. Chapter 7: Sample size issues for conjoint analysis studies. In "Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research" (2010, 2019) Sequim: Sawtooth Software Technical Paper; 1998. (<https://sawtoothsoftware.com/resources/technical-papers/sample-size-issues-for-conjoint-analysis-studies>)
36. Parady, G., K.W. Axhausen, 2024. Size matters: the use and misuse of statistical significance in discrete choice models in the transportation academic. *Transportation*, 51:2393–2425 <https://doi.org/10.1007/s11116-023-10423-y>
37. Pavlou, M., G. Ambler, C. Qu, S.R. Seaman, I.R. White, R.Z. Omar, 2024. An evaluation of sample size requirements for developing risk prediction models with binary outcomes. *BMC Medical Research Methodology*, Vol. 24, 146.
<https://doi.org/10.1186/s12874-024-02268-5>
38. Pate, A., R.D. Riley, G.S. Collins, M. van Smeden, B. Van Calster, J. Ensor, G.P. Martin, 2023. Minimum sample size for developing a multivariable prediction model using multinomial logistic regression. *Statistical Methods in Medical Research*. Vol. 32(3), 555–571.
39. Rainey, C., K. McCaskey, 2021. Estimating logit models with small samples. *Political Science Research and Methods*, 9, 549–564. doi:10.1017/psrm.2021.9
40. Riley, R.D., L. Archer, K.I.E. Snell, J. Ensor, P. Dhiman, G.P. Martin, L.J. Bonnett, G.S. Collins, 2024b. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *British Medical Journal*, Vol. 384, e074820.
<http://dx.doi.org/10.1136/>
41. Riley, R.D., T.P.A. Debray, G.S. Collins, L. Archer, J. Ensor, M. van Smeden, K.I.E. Snell, 2021a. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Statistics in Medicine* Vol. 40, No. 19, pp. 4230-4251.
<https://doi.org/10.1002/sim.9025>

42. Riley, R.D., J. Ensor, K.I.E. Snell, L. Archer, R. Whittle, P. Dhiman, J. Alderman, X. Liu, L. Kirton, J. Manson-Whitton, M. van Smeden, K.G. Moons, K. Nirantharakumar, J.-B. Cazier, A.K. Denniston, B. Van Calster, G.S. Collins, 2025. Importance of sample size on the quality and utility of AI-based prediction models for healthcare. *Lancet Digit Health*, Vol. 7, 100857. <https://doi.org/10.1016/j.landig.2025.01.013>
43. Riley, R.D., K.I.E. Snell, L. Archer, J. Ensor, T.P.A. Debray, B. Van Calster, M. van Smeden, G.S. Collins, 2024b. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. *British Medical Journal*, Vol. 384, e074821. doi: <http://dx.doi.org/10.1136/bmj-2023-074821>
44. Riley, R.D., K.I.E. Snell, J. Ensor, D.L. Burke, F.E. Harrell Jr, K.G.M. Moons, G.S. Collins, 2019a. Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. *Statistics in Medicine*. Vol. 38, No.7, 1262-127. <https://doi.org/10.1002/sim.7993>
45. Riley, R.D., K.I.E. Snell, J. Ensor, D.L. Burke, F.E. Harrell Jr., K.G.M. Moons, G.S. Collins, 2019b. Minimum sample size for developing a multivariable prediction model: PART II – binary and time-to-event outcomes. *Statistics in Medicine*, Vol. 38, 1276–1296. (minor correction: Correction to: Minimum sample size for developing a multivariable prediction model: Part II-binary and time-to-event outcomes by Riley RD, Snell KI, Ensor J, et al. <https://doi.org/10.1002/sim.8409>)
46. Riley, R.D., K. I.E. Snell, G.P. Martin, R. Whittle, L. Archer, M. Sperrin, G.S. Collins, 2021b. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *Journal of Clinical Epidemiology*, Vol. 132, 88-96. <https://doi.org/10.1016/j.jclinepi.2020.12.005>
47. Rose, J.M., M.C.J. Bliemer, 2013. Sample size requirements for stated choice experiments. *Transportation*. Vol. 40, 1021–1041 DOI 10.1007/s11116-013-9451-z
48. Rose, J.M., M.C.J. Bliemer, D.A. Hensher, A.T. Collins, 2008. Designing efficient stated choice experiments in the presence of reference alternatives. *Transportation Research Part B*, 42, 395–406. doi:10.1016/j.trb.2007.09.002.
49. Sadatsafavi, M., P. Gustafson, S. Setayeshgar, L. Wynants, R.D. Riley, 2026. Bayesian Sample Size Calculations for External Validation Studies of Risk Prediction Models. *Statistics in Medicine*, Vol 45, e70389. <https://doi.org/10.1002/sim.70389>
50. Snell, K.E., L. Archer, J. Ensor, L.J. Bonnett, T.P.A. Debray, B. Phillips, G.S. Collins, R.D. Riley, 2021. External validation of clinical prediction models: simulation-based

sample size calculations were more reliable than rules-of-thumb. *Journal of Clinical Epidemiology*, Vol. 135, 79–89. <https://doi.org/10.1016/j.jclinepi.2021.02.011>

51. Sriwastava, A., P. Reichert, 2023. Reducing sample size requirements by extending discrete choice experiments to indifference elicitation. *Journal of Choice Modelling*, Vol. 48, 100426. <https://doi.org/10.1016/j.jocm.2023.100426>
52. Steyerberg, E.W., 2009. *Clinical Prediction Models: A Practical Approach to Development, Application and Updating*. Series of Statistics for Biology and Health. Springer Science+Business Media, New York, NY.
53. Steyerberg, E.W., M.J.C. Eijkemans, F.E. Harrell Jr, J.D.F. Habbema, 2000. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics In Medicine*, Vol. 19, No. 8, 1059-1079. [https://doi.org/10.1002/\(sici\)1097-0258\(20000430\)19:8%3C1059::aid-sim412%3E3.0.co;2-0](https://doi.org/10.1002/(sici)1097-0258(20000430)19:8%3C1059::aid-sim412%3E3.0.co;2-0) 1 (<https://pubmed.ncbi.nlm.nih.gov/10790680/>)
54. Tervonen, T., F. Pignatti, D. Postmus, 2019. From Individual to Population Preferences: Comparison of Discrete Choice and Dirichlet Models for Treatment Benefit-Risk Tradeoffs. *Medical Decision Making*, Vol. 39 (7), 879–885. DOI: 10.1177/0272989X19873630
55. Tian, Y., H. Rusinek, A.V. Masurkar, Y. Feng, 2024. ℓ_1 -Penalized Multinomial Regression: Estimation, Inference, and Prediction, With an Application to Risk Factor Identification for Different Dementia Subtypes. *Statistics in Medicine*, Vol. 43, 5711–5747. <https://doi.org/10.1002/sim.10263>
56. Van Calster, B., M. van Smeden, B. De Cock, E.W. Steyerberg, 2020. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. *Statistical Methods in Medical Research*, Vol. 29, No. 11. 3166–3178. DOI: 10.1177/0962280220921415
57. Van Hoorde, K., Y. Vergouwe, D. Timmerman, S. Van Huffel, E.W. Steyerberg and B. Van Calster, 2014. Assessing calibration of multinomial risk prediction models. *Statistics in Medicine*, Vol. 33, Issue 15, pp. 2585–2596.
58. van Smeden, M., J.A., H. de Groot, K.G.M. Moons, G.S. Collins, D.G. Altman, M.J.C. Eijkemans, J.B. Reitsma, 2016. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*. Vol. 16, No. 163. DOI 10.1186/s12874-016-0267-3

59. Wang, S., B. Mo, Y. Zheng, S. Hess, J. Zhao, 2024. Comparing hundreds of machine learning and discrete choice models for travel demand modeling: An empirical benchmark. *Transportation Research Part B*, Vol. 190, 103061
<https://doi.org/10.1016/j.trb.2024.103061>
60. Wang, Y., A.E. Boyd, L. Rountree, Y. Ren, K. Nyhan, R. Nagar, J. Higginbottom, M.L. Ranney, H. Parikh, B. Mukherjee, 2026. Ten Core Concepts for Ensuring Data Equity in Public Health. *JAMA Health Forum*, Vol. 7 (1), e256031.
doi:10.1001/jamahealthforum.2025.6031.
61. Whittle, R., J. Ensor, L. Archer, G.S. Collins, P. Dhiman, A. Denniston, J. Alderman, A. Legha, M. van Smeden, K.G. Moons, J.-B. Cazier, R.D. Riley, K.I.E. Snell, 2025. Extended sample size calculations for evaluation of prediction models using a threshold for classification. *BMC Medical Research Methodology*, Vol. 25, No. 170.
<https://doi.org/10.1186/s12874-025-02592-4>
62. Zahid, F.M., C. Heumann, 2012. Response shrinkage estimation in multinomial logit models. *Journal of Statistical Planning and Inference*. 142, 95–109.
dx.doi.org/10.1016/j.jspi.2011.06.027
63. Zaloumis, S.G., M. Rajasekhar, J. A. Simpson, 2025. How to use learning curves to evaluate the sample size for malaria prediction models developed using machine learning algorithms. *Malaria Journal*, 24, 242 <https://doi.org/10.1186/s12936-025-05479-3>
64. Zeng, M., M. Zhong, J.D. Hunt, 2018. Analysis of the impact of Sample Size, attribute variance and within-sample choice distribution on the estimation accuracy of multinomial logit models using simulated data. *J Syst Sci Syst Eng*, 27(6):771-789
<https://doi.org/10.1007/s11518-018-5359-7>. (A previous version of the paper above has also been presented at ‘The 3rd International Conference on Transportation Information and Safety, June 25 – June 28, 2015, Wuhan, P. R. China under the title “Effect of Within-Sample Choice Distribution and Sample Size on the Estimation Accuracy of Logit Model” 978-1 4799-8694-1/11.)
65. Zhong, J., X. Liu, J. Lu, J. Yang, G. Zhang, et al., 2025. Overlooked and underpowered: a meta-research addressing sample size in radiomics prediction models for binary outcomes. *European Radiology*, Vol. 35, 1146–1156.
<https://doi.org/10.1007/s00330-024 11331-0>

Papers with Excerpts

Note: in this section, excerpts from the papers appear in italic below each reference. Those excerpts are taken word-for-word and the formatting (e.g., reference styles used in the paper) was not changed. The underlined text refers to comments I made or correspondence I had with authors. A link to the website or DOI number is included wherever possible.

Koppelman, F.S., C. Chu, 1983. Effect of sample size on disaggregate choice model estimation and prediction. Transportation Research Record No.944, 60-69.

<https://onlinepubs.trb.org/Onlinepubs/trr/1983/944/944-009.pdf>

Old study that is still relevant today.

Samples on the order of 1,000 to 2,000 observations may be needed for estimation of relatively simple disaggregate choice models. Although some reduction in this requirement may be obtained by Improved sample design, it is unlikely that the final sample requirements can be reduced to less than 1,000 observations.

Assele, S.Y., M. Meulders, M. Vandebroek, 2023. Sample size selection for discrete choice experiments using design features. Journal of Choice Modelling, 49, 100436. DOI: 10.1016/j.jocm.2023.100436.

There have been several attempts to determine the required sample size for DCE studies using one or more of the design features. The earliest attempt by Orme (1998) provided a formula to approximate the sample size needed based on the number of choice tasks, the number of alternatives, and the largest number of levels for any of the attributes. This formula was revised by Tang et al. (2006) to better represent the heterogeneity of the sample and the complexity of the model. Recently, Yang et al. (2015) have used regression models to calculate the sample size for DCE health studies using several design features and design efficiency. Although the rules of thumb are not accurate, they have been commonly used in several fields. For example, a review by de Bekker-Grob et al. (2015) shows that of 69 healthcare DCE studies that were conducted since 2012, 13% used one or more of the rules of thumb proposed by Orme (1998), by Pearmain and Gleave (1991), or by Lancsar and Louviere (2008) to determine the sample size.

Fig. 4 presents the true sample size calculated with and without taking design complexity into account. As expected, we need larger sample sizes in the presence of choice set complexity to maintain the same power level. As an illustration, Fig. 5 shows the RPE of sample sizes estimated using the $QR(\tau = 0.7)$ model when the true sample size is calculated with and without taking the choice set complexity into account. As expected, the underestimation gets larger as the true sample sizes became larger.

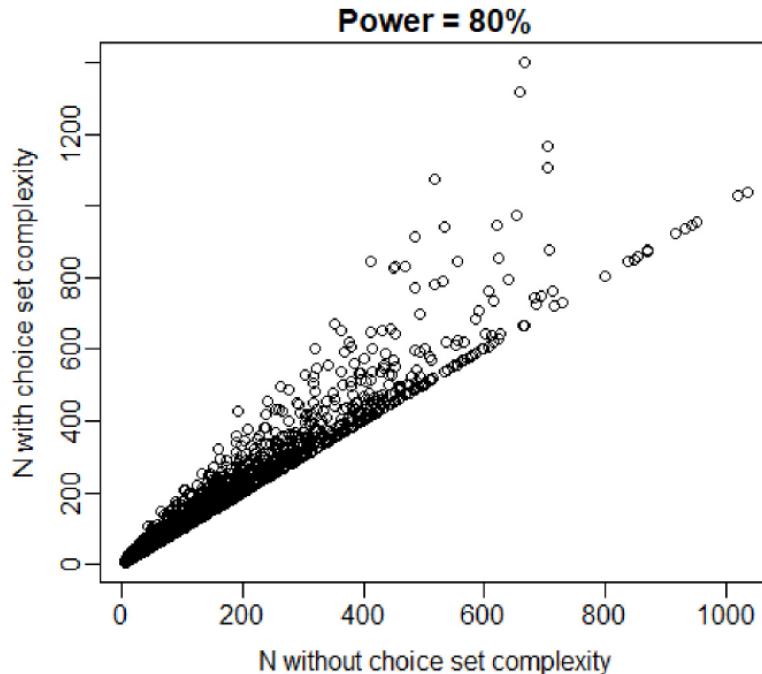


Fig. 4. Scatter plot of the true required sample size computed with and without taking choice set complexity into account.

Last but not least, it is very important to note that the study was restricted to the MNL model whereas the mixed logit model, latent class model, nested logit model, or other extensions of the MNL model will generally require a larger sample.

Bujang MA, Sa'at N, Tg Abu Bakar Sidik TMI, Lim CJ., 2023. Sample size guidelines for logistic regression from observational studies with large population: emphasis on the accuracy between statistics and parameters based on real life clinical data. Malays J Med Sci.; 25(4):122–130. <https://doi.org/10.21315/mjms2018.25.4.12>

Methods: We estimated the minimum sample size required based on evaluation from real clinical data to evaluate the accuracy between statistics derived and the actual parameters. Nagelkerke r-squared and coefficients derived were compared with their respective parameters.

Results: With a minimum sample size of 500, results showed that the differences between the sample estimates and the population was sufficiently small. Based on an audit from a medium size of population, the differences were within ± 0.5 for coefficients and ± 0.02 for Nagelkerke r-squared. Meanwhile for large population, the differences are within ± 1.0 for coefficients and ± 0.02 for Nagelkerke r-squared.

Conclusions: For observational studies with large population size that involve logistic regression in the analysis, taking a minimum sample size of 500 is necessary to derive the statistics that represent the parameters. The other recommended rules of thumb are EPV of 50 and formula; $n = 100 + 50i$ where i refers to number of independent variables in the final model.

de Bekker-Grob, E.W., B. Donkers, M.F. Jonker, E.A. Stolk, 2015. Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide. Patient, 8, 373–384.

The use of sample size calculations for healthcare-related DCE studies is largely lacking. We have shown how sample size calculations can be conducted for DCEs when researchers are interested in testing whether a particular attribute (level) affects the choices that patients or physicians make. Such sample size calculations should be executed far more often than is currently the case in healthcare, as under-powered studies may lead to false insights and incorrect decisions for policy makers.

de Jong, V.M.T., M.J.C. Eijkemans, B. van Calster, D. Timmerman, K.G.M. Moons, E.W. Steyerberg, and M. van Smeden., 2019. Sample size considerations and predictive performance of multinomial logistic prediction models. Statistics in Medicine, Vol. 38, Issue 9, pp. 1601-1619. DOI: 10.1002/sim.8063

(This paper was cited more than 112 times (in SCOPUS – March 23rd, 2026) since 2019 and most of the citations were related to explaining how the paper under discussion had methodological issues related to the sample size used in the paper and biased/erroneous results.)

We present a full-factorial simulation study to examine the predictive performance of MLR models in relation to the relative size of outcome categories, number of predictors and the number of events per variable. It is shown that MLR estimated by Maximum Likelihood yields overfitted prediction models in small to medium sized data. In most cases, the calibration and overall predictive performance of the multinomial prediction model is improved by using penalized MLR. Our simulation study also highlights the importance of events per variable in the multinomial context as well as the total sample size. As expected, our study demonstrates the need for optimism correction of the predictive performance measures when developing the multinomial logistic prediction model. We recommend the use of penalized MLR when prediction models are developed in small data sets or in medium sized data sets with a small total sample size (i.e., when the sizes of the outcome categories are balanced). Finally, we present a case study in which we illustrate the development and validation of penalized and unpenalized multinomial prediction models for predicting malignancy of ovarian cancer.

Pate, A., R.D. Riley, G.S. Collins, M. van Smeden, B. Van Calster, J. Ensor, G.P. Martin, 2023. Minimum sample size for developing a multivariable prediction model using multinomial logistic regression. Statistical Methods in Medical Research. 2023, Vol. 32(3), 555–571.

(This paper was cited more than 128 times (in SCOPUS – March 23rd, 2026) since 2023 and most of the citations were related to explaining how the paper under discussion had methodological issues related to the sample size used in the paper and biased/erroneous results.)

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.

Månsson, K., G. Shukur, B.M. Golam Kibria, 2018. Performance of some ridge regression estimators for the multinomial logit model, Communications in Statistics - Theory and Methods, 47:12, 2795-2804, DOI: 10.1080/03610926.2013.784996

Secondly, as the sample size becomes larger the estimated MSE decreases for both ML and MNLRR. The percentage of times ML is better than MNLRR increases with the sample size in general, which indicates that the gain of applying MNLRR is greater when the sample size is small. Finally, as the number of explanatory variables increases the MSE also becomes higher and the proportion of times ML outperforms MNLRR decreases. Hence, the negative effect of increasing the number of explanatory variables is higher for ML than MNLRR. The result from the simulation study clearly shows that MNLRR outperforms ML. Note: ML: multinomial logit MNLRR: multinomial logit ridge regression.

Riley, R.D., K.I. Snell, J. Ensor, D.L. Burke, F.E. Harrell Jr., K.G.M. Moons, G.S. Collins, 2019. Minimum sample size for developing a multivariable prediction model: PART II – binary and time-to-event outcomes. Statistics in Medicine. 38: 1276–1296.

(This paper was cited more than 827 times (in SCOPUS – March 23rd, 2026) since 2019 and a large proportion of the citations were related to explaining how the paper under discussion had methodological issues related to the sample size used in the paper and biased/erroneous results.)

We propose that the minimum values of n and E (and subsequently the minimum number of events per predictor parameter, EPP) should be calculated to meet the following three 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 model's apparent and adjusted Nagelkerke's R^2 , and (iii) precise estimation of the overall risk in the population. Criteria (i) and (ii) aim to reduce overfitting conditional on a chosen p , and require prespecification of the model's anticipated Cox-Snell R^2 , which we show can be obtained from previous studies. The values of n and E that meet all three criteria provides the minimum sample size required for model development. Upon application of our approach, a new diagnostic model for Chagas disease requires an EPP of at least 4.8 and a new prognostic model for recurrent venous thromboembolism requires an EPP of at least 23. This reinforces why rules of thumb (e.g., 10 EPP) should be avoided. Researchers might additionally ensure the sample size gives precise estimates of key predictor effects; this is especially important when key categorical predictors have few events in some categories, as this may substantially increase the numbers required.

Farghali, R.A., M. Qasim, B.M. Golam Kibria, M.R. Abonazel (2023) Generalized two-parameter estimators in the multinomial logit regression model: methods, simulation and application, Communications in Statistics - Simulation and Computation, 52:7, 3327-3342, DOI: 10.1080/03610918.2021.1934023

Namely, when the sample size n increases, the estimated SMSE values decrease. It is obvious from tables and figures that by increasing the sample size affects positively on the performance of all estimators (including MLE).

Van Hoorde, K., Y. Vergouwe, D. Timmerman, S. Van Huffel, E.W. Steyerberg and B. Van Calster, 2014. Assessing calibration of multinomial risk prediction models. Statistics in Medicine, Vol. 33, Issue 15, pp. 2585–2596.

We propose a multinomial logistic recalibration framework that involves an MLR fit where Y is predicted using the $k-1$ linear predictors from the prediction model. A non-parametric alternative may use vector splines for the effects of the linear predictors. The parametric and non-parametric frameworks can be used to generate multinomial calibration plots. Further, the parametric framework can be used for the estimation and statistical testing of calibration intercepts and slopes. (This paper explains how to apply the Calibrating slope method to assess whether or not the model is overfitted).

Abonazel, M.R., I. Dawoud, M. Naji Al-Ghamdi, R.A. Farghali, 2025. Developing the generalized Dawoud-Kibria estimator for the multinomial logistic model: Simulation study and application. Scientific African 29, e02803. <https://doi.org/10.1016/j.sciaf.2025.e02803>

At this stage, it is clear that the ML estimator is highly unreliable for moderate sample sizes ($n \leq 500$) when multicollinearity is present. While some biased estimators like GMLT and GMRR perform well in specific cases, GMDK consistently provides the lowest MSE across all conditions.

Orme, B. 2019. Chapter 7: Sample size issues for conjoint analysis studies. In "Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research" (2010, 2019) Sequim: Sawtooth Software Technical Paper; 1998. (<https://sawtoothsoftware.com/resources/technical-papers/sample-size-issues-for-conjoint-analysis-studies>)

The recommendations below assume infinite or very large populations. They are based on the theories above and our observations of common practices in the market research community:

Sample sizes for conjoint studies generally range from about 150 to 1,200 respondents. If the purpose of your research is to compare groups of respondents and detect significant differences, you should use a large enough sample size to accommodate a minimum of about 200 per group. Therefore, if you are conducting a segmentation study and plan to divide respondents into as many as four groups (i.e., through cluster analysis) it would be wise to include, at a minimum, $4 \times 200 = 800$ respondents. This, of course, assumes your final group sizes will be about equal, so one would usually want more data. The stronger segmentation studies include about 800 or more respondents.

For robust quantitative research where one does not intend to compare subgroups, I would recommend at least 300 respondents. For investigational work and developing hypotheses about a market, between thirty and sixty respondents may do.

Martin, G.P., M. Sperrin, K.I.E. Snell, I. Buchan, R.D. Riley, 2021. Clinical prediction models to predict the risk of multiple binary outcomes: a comparison of approaches. Statistics in Medicine. 2021;40:498–517. DOI: 10.1002/sim.8787

Third, we only considered case studies and simulations with large sample sizes; therefore, further research is needed to explore the concepts of this paper in settings where overfitting might be a concern (e.g., penalization).²⁸-> [Riley et al. \(2019\)](#) above.

Zahid, F.M., C. Heumann, 2012. Response shrinkage estimation in multinomial logit models. Journal of Statistical Planning and Inference. 142, 95–109. dx.doi.org/10.1016/j.jspi.2011.06.027

The proposed method handles not only the problem of separation in multinomial logit models but estimates also exist when the number of covariates is large relative to the sample size.

Fuetterer, C., M. Nalenz, T. Augustin, R.M. Pfeiffer, 2025. A powerful penalized multinomial logistic regression approach. Computational Statistics <https://doi.org/10.1007/s00180-025-01635-0>

When the number p of (correlated) predictors was much larger than the available sample size N , DPan had the highest true positive rate while maintaining low false positive rates for all simulation settings. Similarly, when $p < N$, DPan had high true positive rates and the lowest false positive rates of all methods studied.

Biscadi, S., J. Orprecio, M. Brock Fenton, A. Tsoar, J.M. Ratcliff, 2004. Data, sample sizes and statistics affect the recognition of species of bats by their echolocation calls. Acta Chiropterologica, 6(2): 347–363. <https://doi.org/10.3161/001.006.0212>

We document the effects of sample sizes and a priori assignment of calls by species on the outcome of discriminant function analysis (DFA) and multinomial logistic regression (MLR) of features of echolocation calls, and determine which features of calls are most useful for identification... Outcomes of DFA and MLR were affected by both sample sizes (numbers of calls, numbers of sequences) and the subjective approach that researchers take to their data (i.e., categorizing calls or sequences of calls by species).

Wang, S., B. Mo, Y. Zheng, S. Hess, J. Zhao, 2024. Comparing hundreds of machine learning and discrete choice models for travel demand modeling: An empirical benchmark. Transportation Research Part B 190 (2024) 103061 <https://doi.org/10.1016/j.trb.2024.103061>

This paper supports the results of Ye and Lord (2013) for the MNL, where the MNL performance improved from 1,000 observations to 10,000 observations, but remained similar for 100,000 observations.

*Firstly, many ML models, particularly the ensemble methods and deep learning, statistically outperform the DCM family and its individual variants (i.e., multinomial, nested, and mixed logit), thus corroborating with the previous research. However, this study also highlights the crucial role of the contextual factors (i.e., data sources, inputs and choice categories), which can explain models' predictive performance more effectively than the differences in model types alone. Model performance varies significantly with data sources, **improving with larger sample sizes** and lower dimensional alternative sets.*

Personal communication:

“Dear Dominique:

You are absolutely right - most of the studies don't compare to the MXL models potentially due to computational challenges. Indeed, sample size is a crucial factor. Whenever the sample size is small, some regularization (or penalty, as pointed out by your "Sample size considerations and predictive performance of multinomial logistic prediction models") typically can improve the performance.

Sample size is important because the generalization bound of the predictive performance converges with sample size. I have a relatively theoretical paper discussing sample size - I am attaching the paper here.

Let me know if you have further thoughts!"

Gehringer, C.K., G.P. Martin, K.L. Hyrich, S.M.M. Verstappen, J. Sextone, E.K. Kristianslund, S.A. Provane, T.K. Kvien, J.C. Sergeant, 2024. Developing and externally validating multinomial prediction models for methotrexate treatment outcomes in patients with rheumatoid arthritis: results from an international collaboration. Journal of Clinical Epidemiology 166 (2024) 111239. <https://doi.org/10.1016/j.jclinepi.2023.111239>

A sample size calculation for the development of multinomial prediction models was carried out [32]. The calculation suggested that a maximum of eight candidate predictor parameters could be included given the fixed sample size (further detail in Supplementary material)... We used the proposed sample size calculation for external validation of a CPM [46] to obtain the minimum requirements for precise estimates of calibration (Observed/Expected, c-slope) and discrimination (c-statistic). As this calculation was developed for models with a binary outcome, we calculated the minimum required sample size for each submodel (outcome pair in the multinomial model) and chose the larger value (results in Supplementary material).

Model development sample size calculation

The sample size available for model development was fixed at 1,632 due to using existing data. We therefore conducted a calculation to determine how many predictor parameters can be included during model development given the fixed sample size. The prevalence of each outcome category: 1) no LDA = 756, 2) LDA = 730, and 3) discontinuation due to AEs = 146, was included in the calculation. A conservative approach was used. For Criterion 1, the shrinkage factor was set to 0.9 and for Criterion 2, the expect value of the maximum Cox-Snell R² was set to 15%. From this, we established the total number of predictor parameters that could be included for the fixed available sample size. In our example, this meant that the number of predictors was changed until the total sample size given by the calculation was less than or equal to 1,632. The calculation suggested that eight predictor parameters could be used during model development.

Cao, Q., Z. Jiang, Z. Wang, L. Wee, A. Dekker, Z. Zhang, J. Zhu, 2025. Minimum sample size calculation for radiomics-based binary outcome prediction models: Theoretical framework and practical example. Radiotherapy and Oncology, Vol. 212, 111134. <https://doi.org/10.1016/j.radonc.2025.111134>

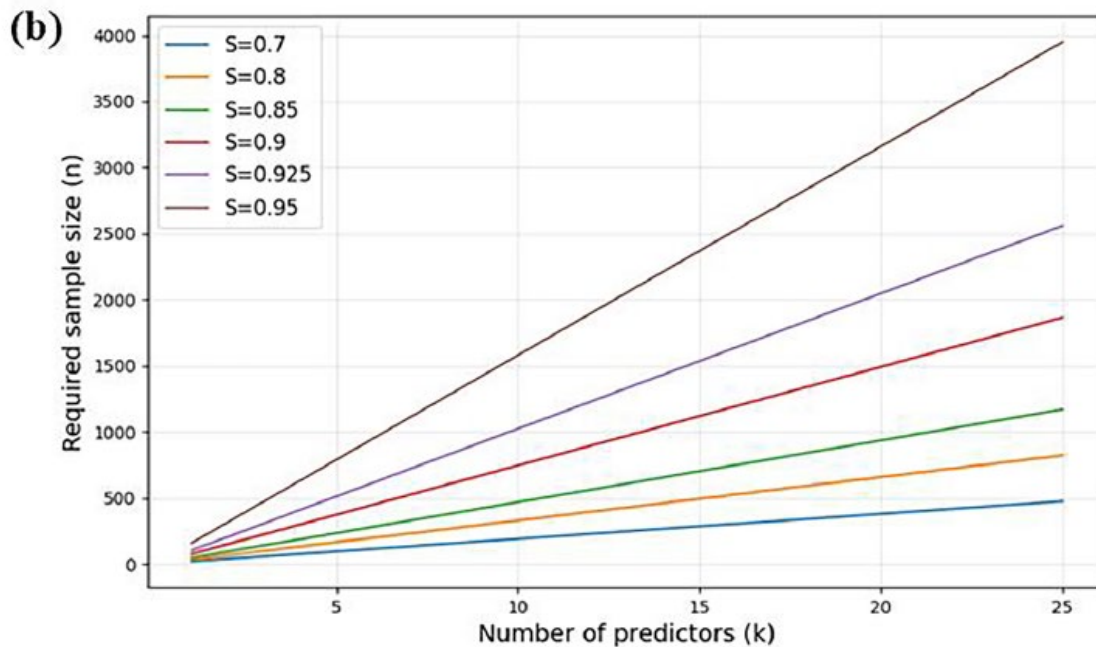
Background and purpose: Determining the appropriate sample size for developing robust radiomics-based binary outcome prediction models and identifying the maximum number of predictors safely allowable within a fixed dataset size remain critical yet challenging tasks. This

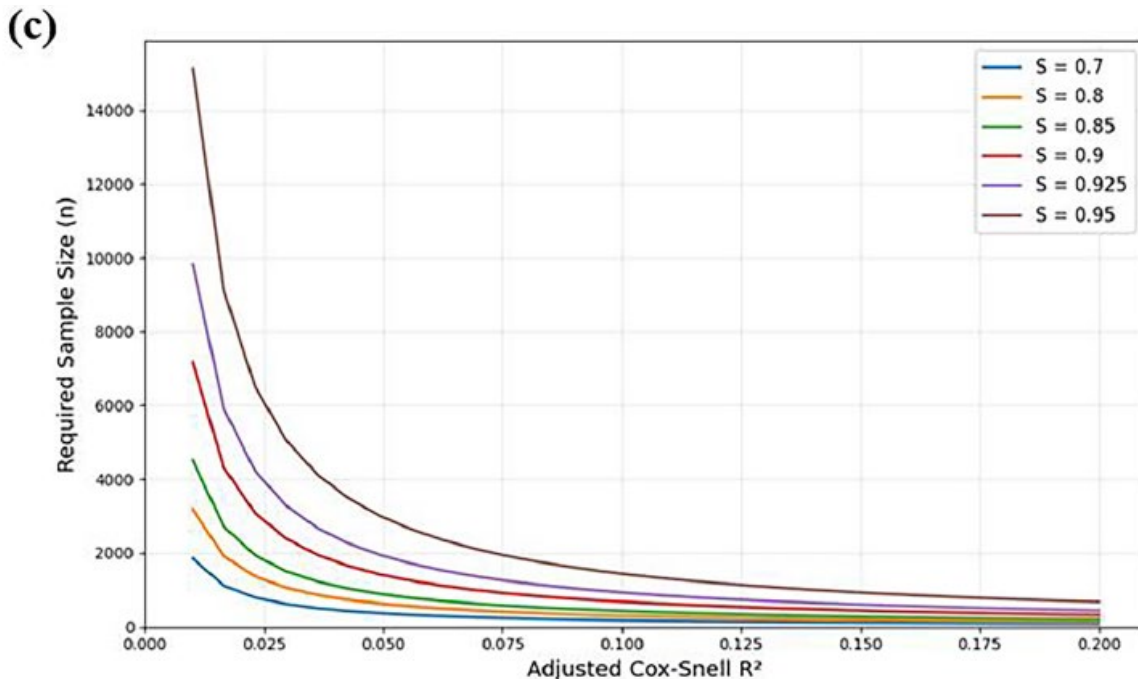
study aims to propose and demonstrate a structured method for addressing these issues, enhancing methodological rigor and practicality in radiomics research.

Materials and methods: We introduce a comprehensive sample size calculation framework for binary outcome prediction models in radiomic studies. The proposed approach integrates three key criteria: (1) maintaining a global shrinkage factor (S) ≥ 0.9 to control model overfitting, (2) ensuring a minimal absolute difference between apparent and adjusted performance metrics, and (3) precisely estimating the overall outcome risk. Additionally, we develop an accessible online calculation tool enabling researchers to efficiently determine either the minimum sample size or the maximum number of predictors permissible, based on clearly defined statistical parameters.

Results: The presented method systematically addresses model overfitting by integrating a global shrinkage factor into the calculation, providing robust estimates compared with traditional heuristic approaches (“rules of thumb”). Practical examples demonstrate that this structured method effectively balances predictive accuracy and generalizability, while the online tool provides researchers with a user-friendly platform to perform the necessary calculations.

Conclusion: Clear justification of sample size decisions is essential for developing reliable predictive models in radiomics research. By adopting a structured and rigorous calculation method, researchers can effectively minimize overfitting, ensure accurate risk estimation, and substantially enhance the reliability and validity of their predictive models.





Elayan, H., M. Sperrin, G.P. Martin, N. Peek, F. Braunschweig, J. Faxén, J. Alfredsson. D.A. Jenkins, 2025. Correcting for case-mix shift when developing clinical prediction models. *BMC Medical Research Methodology*, Vol. 25, 186. <https://doi.org/10.1186/s12874-025-02621-2>

Furthermore, the minimum sample size criteria was calculated using the method of Riley et al. [22], where we assumed an event prevalence of 0.0021, a target shrinkage factor of 0.9, number of predictors of 7, and c-statistic of 0.75 [40], this gave a required sample size of 33,123. This minimum sample size criteria was met during the models' development.

Gehring, C.K., G.P. Martin, B. Van Calster, K.L. Hyrich, S.M.M. Verstappen, J.C. Sergeant, 2024. How to develop, validate, and update clinical prediction models using multinomial logistic regression. *Journal of Clinical Epidemiology* 174 (2024) 111481. <https://doi.org/10.1016/j.jclinepi.2024.111481>

Sample size guidance for developing CPMs for continuous, binary and time-to-event models has been developed in the past few years [56e58], with extensions for the development of an MPM recently proposed [29]. Crucially, these sample size calculations help to minimize the level of overfitting of the CPM, and they ensure that there is sufficient sample size to precisely estimate key model parameters (such as the model intercept). A calculation should be performed before the development of an MPM, to determine the maximum number of predictor parameters relative to the number of participants, outcome prevalence and expected predictive performance. The total sample size required for an MPM depends on the number of outcome categories, and more outcome categories require larger sample sizes [29].

Dekker, T., P. Bansal, J. Huo, 2025. Revisiting McFadden's correction factor for sampling of alternatives in multinomial logit and mixed multinomial logit models. *Transportation Research Part B*, 103-129. <https://doi.org/10.1016/j.trb.2024.103129>

As sample sizes become sufficiently large consistent point estimates for MNL can be obtained as per McFadden's original proof. McFadden's correction factor can therefore effectively be applied in the context of Bayesian MNL models. We extend these results to the context of mixed multinomial logit models (MMNL) by using the property of data augmentation in Bayesian estimation. McFadden's correction factor minimises the expected information loss with respect to the augmented individual-level parameters, and in turn also for the population parameters characterising the shape and location of the mixing density in MMNL. Again, the results apply to finite and large samples and most importantly circumvent the need for additional correction factors previously identified for estimating MMNL models using maximum simulated likelihood.

As the sample size increases, the randomness of uniform conditioning increases the probability that all relevant utility differences are studied in a balanced way across the sample which is desirable to obtain consistent parameter estimates.... ...As the sample size increases it can be shown that the Bayesian point estimate converges to the (consistent) point estimate obtained using classical maximum likelihood estimation... .. As the sample size decreases two effects occur. First, posterior and classical standard errors on the parameter estimates increase because there is less information in the data regarding β by default. Second, sampling of alternatives is likely to generate some additional degree of error because we sample fewer chosen alternatives (from the population) and choice sets D_n (when applying sampling of alternatives) across the sample. Only when the sample size becomes sufficiently large, the actual information loss with respect to β (in the parameter estimate and in the posterior) will converge in probability to the expected information loss (as per McFadden(1978)'s proof in Appendix A)... ...When the sample size increases either by increasing the number of respondents or the number of choices per respondent we generally see a reduction in the variability of the parameter estimates, their bias, increases in the average coverage probability and reductions in the average standard error of the parameter estimates.

Zhong, J., X. Liu, J. Lu, J. Yang, G. Zhang, et al., 2025. Overlooked and underpowered: a meta-research addressing sample size in radiomics prediction models for binary outcomes. European Radiology (2025) 35:1146–1156, <https://doi.org/10.1007/s00330-024-11331-0>

To investigate how studies determine the sample size when developing radiomics prediction models for binary outcomes, and whether the sample size meets the estimates obtained by using established criteria... ...Radiomics studies are often designed without sample size justification, whose sample size may be too small to avoid overfitting. Sample size justification is encouraged when developing a radiomics model.

Conclusion Radiomics studies are often designed without sample size justification, whose sample size may be too small to avoid overfitting. Sample size justification is encouraged when developing a radiomics model.

Key Points

Question Sample size justification is critical to help minimize overfitting in developing a radiomics model, but is overlooked and underpowered in radiomics research.

Martin, G.P., R.D. Riley, J. Ensor, S.W. Grant, 2025. Statistical primer: sample size considerations for developing and validating clinical prediction models. European Journal of Cardio-thoracic Surgery; doi:10.1093/ejcts/ezaf142.

It is crucial to ensure that the sample size of the data used to develop or validate a clinical prediction model is large enough. If the data are inadequate, developed models can be unstable and estimates of predictive performance imprecise. This can lead to models that are unfit or even harmful for clinical practice. Recently, there have been a series of sample size formulae developed to estimate the minimum required sample size for prediction model development or external validation. The aim of this statistical primer is to provide an overview of these criteria, describe what information is required to make the calculations and illustrate their implementation through worked examples.

... When developing a CPM, the overall goal is to fit a statistical model or train a machine learning algorithm that provides accurate risk predictions in new patients from the population where the model is intended to be used in practice. The new sample size guidance identifies the minimum observations needed to achieve this.

There are 2 main use cases for the proposed sample size criteria when planning to develop a CPM. First, they may be used to facilitate prospective data collection by determining how many observations need to be collected to support CPM development. Second, the calculations can be used when the sample size available is fixed (as is the case with most retrospective studies) to determine if the dataset available is sufficient. In either case, the criteria can also be used to guide changes to the study design, as required. For example, reducing the number of candidate predictor parameters considered for the CPM until the minimum sample size matches the available sample size.

Fundamentally, the sample size criteria [15, 17–20] aim to ensure the sample size will precisely estimate the overall outcome event proportion (risk) in the population. If the sample size is not even large enough to estimate the overall risk precisely at the population level, then it is futile to consider estimating individual-level risks. Moreover, the criteria aim to minimize overfitting of the model or algorithm.

Overfitting means that the estimated predictor effects included in the model are too extreme and the model is unlikely to perform reliably outside of the development dataset. Putting this another way, the model is overly complex (in terms of number of parameters) relative to the available data, meaning the model fit too closely matches nuances that are contained within the development data, to the detriment of predictions in new individuals. To help combat overfitting, shrinkage (penalization) methods that aim to reduce extreme predictor effects can be applied during CPM development [25]. One such method is the heuristic shrinkage factor [26], which applies an overall multiplicative factor to reduce the parameter values towards zero (Fig. 1).

However, although these methods are important, they do not overcome the need for developing CPMs using a sufficient sample size, as shrinkage factors are themselves estimated from the data, and small samples lead to instability in shrinkage estimates and ultimately the developed model [27–29].

Worked example 1: 25 variables -> 9,658 observations, 20 variables -> 7,726 observations, 18 variables -> 6,954 observations, 15 variables -> 5,795 observations; Worked example 2: 12,018 observations. Two were for binary logit model.

Zeng, M., M. Zhong, J.D. Hunt, 2018. Analysis of the impact of Sample Size, attribute variance and within-sample choice distribution on the estimation accuracy of multinomial logit models using simulated data. J Syst Sci Syst Eng, 27(6):771-789 <https://doi.org/10.1007/s11518-018-5359-7>

These authors arrived at the same conclusions as Ye and Lord (2014), but recommended a larger sample size for MNL models -> 3,000 to 4,000 observations)

It is found that (1) the estimation accuracy of utility parameters increases as the sample size increases; (2) the utility coefficients can be re-estimated with reasonable accuracy, but the estimates of the ASCs are confronted with much larger errors; (3) as the variances of the alternative attributes increase, the estimation accuracy improves significantly; and (4) as the distribution of chosen choices becomes more balanced across alternatives within sample datasets, the hit-ratio decreases. The results indicate that (a) under a similar setting presented in this paper, a large sample consisting of a few thousand observations (3000 – 4000) may be needed in order to provide reasonable estimates for utility coefficients, particularly for ASCs.

A previous version of the paper above has also been presented at *The 3rd International Conference on Transportation Information and Safety, June 25 – June 28, 2015, Wuhan, P. R. China* under the title **“Effect of Within-Sample Choice Distribution and Sample Size on the Estimation Accuracy of Logit Model”** 978-1-4799-8694-1/11/231.00 © 2015 IEEE

Hippisley-Cox, J., C.A. Coupland, 2025. Development and external validation of prediction algorithms to improve early diagnosis of cancer. Nature Communications, Vol. 16, 3660. <https://doi.org/10.1038/s41467-025-57990-5>

Sample size calculations

We used all eligible individuals to develop and validate the models to maximise the power and generalisability of the results. With the sample sizes in the derivation cohorts for men and women, we had large enough samples to consider up to 100 predictor variables and target a shrinkage value of 0.9, a difference of 0.05 for comparison of R2 values and a margin of error of 0.05 for the estimates of overall probability, with the exception of a small number of more rare cancer pairs where a reduced shrinkage of 0.8 can be targeted with a smaller number of predictor variables considered for inclusion. We performed sample size calculations for both model derivation and validation, as shown in the Supplement. We used Stata (version 18) for analyses.

From the supplement material: https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-025-57990-5/MediaObjects/41467_2025_57990_MOESM1_ESM.pdf

Information on sample size calculations

1. Model derivation sample size

The sample size of the derivation cohort was 3,622,789 for women, and 3,841,718 for men. Using the sample size criteria for developing a multinomial prediction model in Pate et al, we

determined the minimum sample size required based on a maximum of 100 predictor variables and the proportions with each type of cancer. For criterion 1, based on a target shrinkage value, we specified a shrinkage factor of 0.9. For criterion 2, targeting a difference in apparent and adjusted R2 values, we used a maximum Cox-Snell apparent R2 value of 0.22 (derived from equation 28) and a difference of 0.01, and for criterion 3 based on obtaining a precise estimate of overall probability, we used a margin of error value of 0.01.

For criterion 1 the minimum sample size required in women ranged from 88,764 to 1,159,042 for pairwise comparisons of individual cancer types with no cancer. For pairwise comparisons between cancer types the minimum sample sizes required for some of the more rare cancer pairs exceeded the sample size available, but following suggestions in the paper by Pate et al, by reducing shrinkage to 0.8 and the number of predictor variables to 60 for the two rarest cancers (oral and liver) the sample size required was 3,382,393. It should be noted that only a subset of the predictor variables were considered for inclusion in the model for some cancers with fewer relevant symptoms. Similarly in men the minimum sample size required ranged from 134,076 to 926,911 for pairwise comparisons with no cancer. For pairwise comparisons between cancer types, for the two rarest cancers (testicular and liver), the sample size required was 3,227,602 for a shrinkage of 0.8 with 80 predictor variables.

For criterion 2 the minimum sample size required was 112,679 for women, and 115,859 for men.

For criterion 3 the minimum sample size required was 1479 for women, and 1422 for men.

Across these criteria the sample size of the available data met the required minimum sample sizes of 3,382,393 in women and 3,227,602 in men for the specified criteria, allowing for a reduced shrinkage target for the rare cancer pairs where there may be some overfitting.

2. Validation sample size

The sample size of the QResearch validation dataset was 1,277,015 for women, and 1,360,169 for men. The corresponding figures for CPRD were 1,373,006 for women and 1,363,720 for men. We calculated the minimum sample size required for the validation analyses using the criteria in Riley et al. for validation of a prediction model with a binary outcome, as currently there are no specific criteria published for validation of multinomial prediction models. We calculated the minimum sample size required for comparisons of each cancer type against no cancer, and selected the highest of these.

We used the `pmvalsampsize` command in Stata3, with target 95% confidence interval widths of 0.2 for the calibration in the large measure O/E (criterion 1), 0.2 for the calibration slope, and 0.1 for the C statistic based on a value of 0.8.

For criterion 1 the required sample size ranged from 68,271 for breast cancer to 1,281,174 for oral and liver cancer in women, and 116,121 for prostate cancer to 960,785 for testicular cancer in men.

For criterion 2 the required sample size ranged from 11,685 for breast cancer to 471,971 for oral cancer in women, and from 12,649 for prostate cancer to 303,678 for testicular cancer in men.

Snell, K.E.I., L. Archer, J. Ensor, L.J. Bonnett, T.P.A. Debray, B. Phillips, G.S. Collins, R.D. Riley,

External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. *Journal of Clinical Epidemiology*, 135, 79–89.

<https://doi.org/10.1016/j.jclinepi.2021.02.011>

method: Simulation of logistic regression prediction models to investigate factors associated with precision of performance estimates. Then, explanation and illustration of a simulation-based approach to calculate the minimum sample size required to precisely estimate a model's calibration, discrimination and clinical utility.

Results: Precision is affected by the model's linear predictor (LP) distribution, in addition to number of events and total sample size. Sample sizes of 100 (oreven200) events and non-events can give imprecise estimates, especially for calibration. The simulation-based calculation accounts for the LP distribution and (mis)calibration in the validation sample. Application identifies 2430 required participants (531 events) for external validation of a deep vein-thrombosis diagnostic model.

Hamzah, Y.B.W., X.-J. Xie, 2016. Effects of different type of covariates and sample size estimation for multinomial logistic regression model. *Jurnal Teknologi (Sciences & Engineering)* 78:12–3, 155–161

The sample size and distributions of covariate may affect many statistical modeling techniques. This paper investigates the effects of sample size and data distribution on parameter estimates for multinomial logistic regression. A simulation study was conducted for different distributions (symmetric normal, positively skewed, negatively skewed) for the continuous covariates. In addition, we simulate categorical covariates to investigate their effects on parameter estimation for the multinomial logistic regression model. The simulation results show that the effect of skewed and categorical covariate reduces as sample size increases. The parameter estimates for normal distribution covariate apparently are less affected by sample size. For multinomial logistic regression model with a single covariate study, a sample size of at least 300 is required to obtain unbiased estimates when the covariate is positively skewed or is a categorical covariate. A much larger sample size is required when covariates are negatively skewed.

In short, a MNL with a single (skewed or categorical) parameter requires 300 observations. Imagine if the model includes 20 such parameters.

Kim, S., E. Heath, L. Heilbrun, 2017. Sample size determination for logistic regression on a logit-normal distribution. *Statistical Methods in Medical Research* 2017, Vol. 26(3) 1237–1247.
<https://doi.org/10.1177/0962280215572407>

The authors compared different methods for estimating the minimum sample size for different outcomes using simulation. These authors indicated that by transforming the distribution of the logit model to a normal distribution, the minimum sample size requirement can be reduced.

Rainey, C., K. McCaskey, 2021. Estimating logit models with small samples. *Political Science Research and Methods* (2021), 9, 549–564. doi:10.1017/psrm.2021.9

*In small samples, maximum likelihood (ML) estimates of logit model coefficients have substantial bias away from zero. As a solution, we remind political scientists of Firth's (1993, *Biometrika*, 80, 27–38) penalized maximum likelihood (PML) estimator. Prior research has described and used*

PML, especially in the context of separation, but its small sample properties remain under-appreciated. The PML estimator eliminates most of the bias and, perhaps more importantly, greatly reduces the variance of the usual ML estimator. Thus, researchers do not face a bias-variance tradeoff when choosing between the ML and PML estimators—the PML estimator has a smaller bias and a smaller variance. We use Monte Carlo simulations and a re-analysis of George and Epstein (1992, American Political Science Review, 86, 323–337) to show that the PML estimator offers a substantial improvement in small samples (e.g., 50 observations) and noticeable improvement even in larger samples (e.g., 1000 observations).

Zaloumis, S.G., M. Rajasekhar, J. A. Simpson, 2025. How to use learning curves to evaluate the sample size for malaria prediction models developed using machine learning algorithms. Malaria Journal (2025) 24:242 <https://doi.org/10.1186/s12936-025-05479-3>

Riley et al. [1, 4, 5] published recommendations for calculating the sample size needed to develop a clinical prediction model, and present a procedure for continuous, binary and survival (time-to-event) outcomes, where the aim is to minimize the potential for model overfitting and to estimate key parameters evaluating the performance of the prediction model precisely (e.g. the overall outcome proportion and predicted outcome probabilities for new individuals in the case of binary outcomes).

Steyerberg, E.W., M.J.C. Eijkemans, F.E. Harrell Jr, J.D.F. Habbema, 2000. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets STATISTICS IN MEDICINE Statist. Med. 2000; 19:1059}1079 10.1002/(sici)1097-0258(20000430)19:8<1059::aid-sim412>3.0.co;2-0

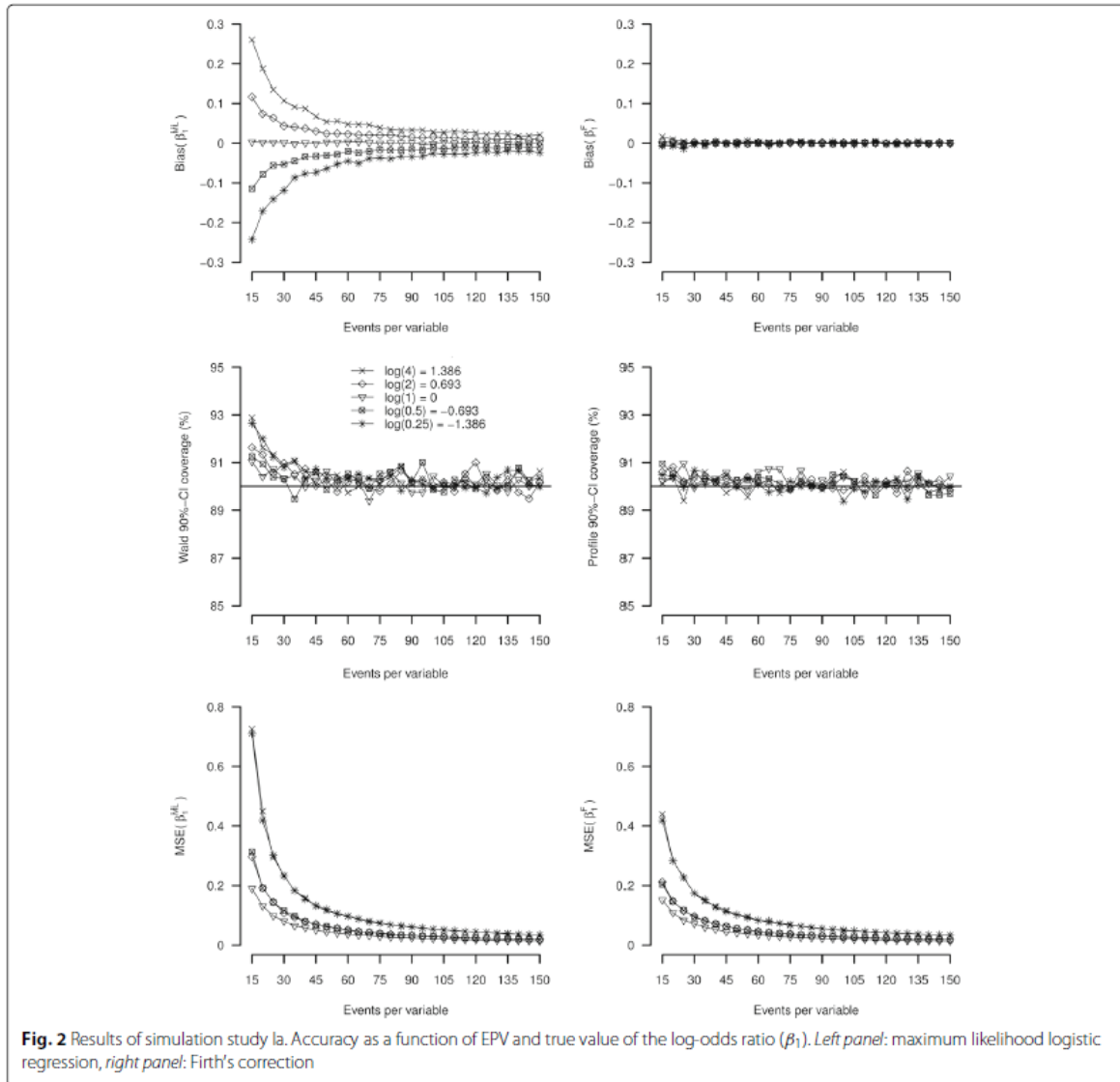
A small data set has in this respect been defined as one with less than 10 events per variable (EPV) [31]. The 1:10 rule is somewhat arbitrary, and we may try to refine this criterion. First, it is evident that the total sample size is also important, in addition to the EPV. This was illustrated by the modelling of 17 predictors in data sets with 62 events, which was much less problematic than modelling eight predictors in data sets with 23 events, although the EPV value was only slightly larger (3.7 or 2.9). Second, we propose two additional critical EPV values: 20 and 50. When the 1:10 rule is violated, the number of parameters to be estimated may in fact be too large for the data under study. A small prespecified model should be fit with shrinkage of the regression coefficients. When the EPV is larger than 10 but smaller than 20, a prespecified model may adequately be fit, but shrinkage is advisable. When the EPV exceeds 20, shrinkage may not be necessary anymore for full models. Criteria for application of stepwise selection with $\alpha=0.05$ are difficult to provide. We performed some additional analyses with 17-predictor models where nine covariables were made randomly associated with the outcome. These analyses indicated that stepwise selection did not improve predictive performance compared with shrunk full models unless EPV exceeded 50 (that is, only in the total training data set). Note that the number of candidate predictors should be considered in this reasoning, not the number of predictors included in the final model. It may hence often be impossible to study a comprehensive set of potential predictors, since this may easily amount to 50 to 100 predictors in prognostic problems. Further research of EPV criteria is indicated.

van Smeden et al., 2016. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Medical Research Methodology 16:163 DOI 10.1186/s12874-016-0267-3

Methods: The current study uses Monte Carlo simulations to evaluate small sample bias, coverage of confidence intervals and mean square error of logit coefficients. Logistic regression models fitted by maximum likelihood and a modified estimation procedure, known as Firth's correction, are compared.

Results: The results show that besides EPV, the problems associated with low EPV depend on other factors such as the total sample size. It is also demonstrated that simulation results can be dominated by even a few simulated data sets for which the prediction of the outcome by the covariates is perfect ('separation'). We reveal that different approaches for identifying and handling separation leads to substantially different simulation results. We further show that Firth's correction can be used to improve the accuracy of regression coefficients and alleviate the problems associated with separation.

Conclusions: The current evidence supporting EPV rules for binary logistic regression is weak. Given our findings, there is an urgent need for new research to provide guidance for supporting sample size considerations for binary logistic regression analysis.



Ogundimu, E.O., D.G. Altman, G.S., Collins, 2016. Adequate sample size for developing prediction models is not simply related to events per variable. *Journal of Clinical Epidemiology*, 76, 175-82.

“Study Design and Setting: We conducted an extended resampling study using a large general-practice data set, comprising over 2 million anonymized patient records, to examine the EPV requirements for prediction models with low-prevalence binary predictors developed using Cox regression. The performance of the models was then evaluated using an independent external validation data set. We investigated both fully specified models and models derived using variable selection.

Results: Our results indicated that an EPV rule of thumb should be data driven and that EPV 20 generally eliminates bias in regression coefficients when many low-prevalence predictors are included in a Cox model.

Conclusion: Higher EPV is needed when low-prevalence predictors are present in a model to eliminate bias in regression coefficients and improve predictive accuracy.”

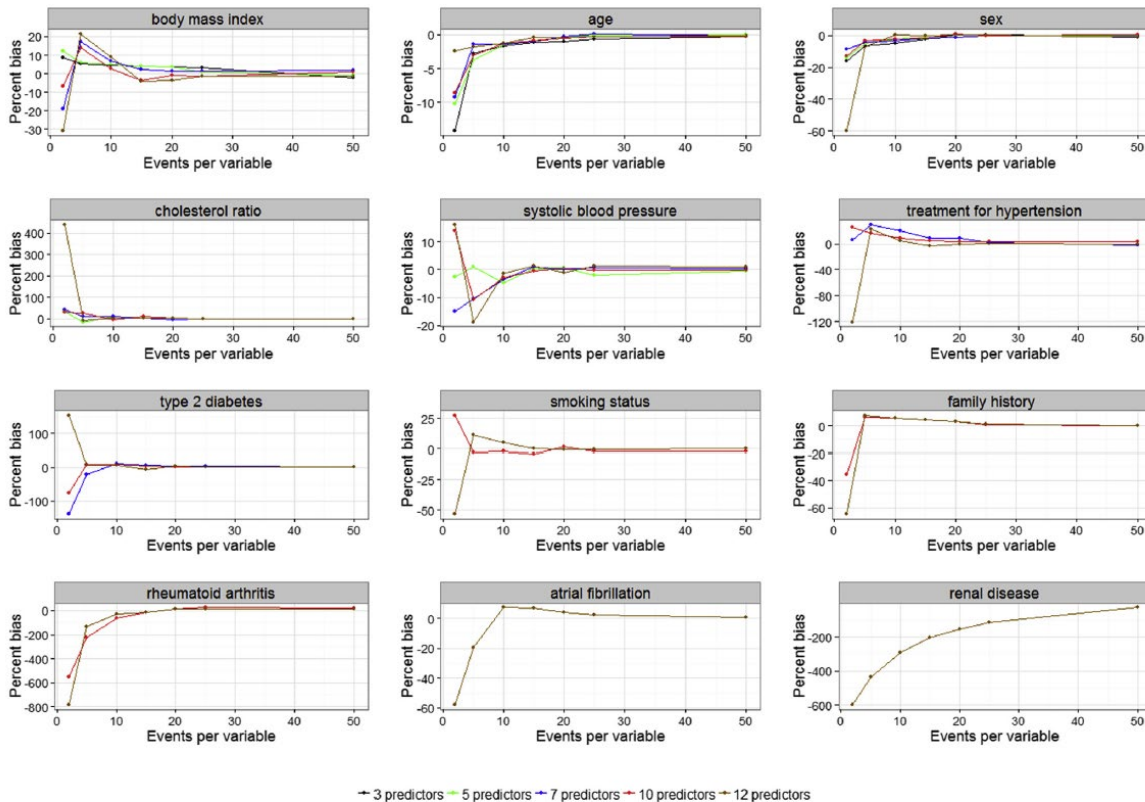


Fig. 1. Number of events per variable and average percent relative bias for the variables in the data set.

Rose, J.M., M.C.J. Bliemer, 2013. Sample size requirements for stated choice experiments. Transportation (2013) 40:1021–1041 DOI 10.1007/s11116-013-9451-z

This paper addresses the issue of how to determine the theoretical minimum sample size for SC studies using the MMNL (mixed MNL) model. In the paper, we argue that the current theory on sample size calculation has not been adequately addressed specifically for SC experiments. From this point, we then go on to demonstrate methods that not only allow for a calculation of the sample size requirements of SC studies, but demonstrate for two different objectives, how analysts may generate SC designs that will minimize the necessary sample size without compromising the reliability of the model results. In developing our arguments, we have assumed the panel version of the MMNL model accounting for the pseudo panel nature of SC experiments.

The required sample size depends on the model type, the number of alternatives, the number of attributes, the number of attribute levels, the attribute level range, the design itself, and the likely parameter estimates. Therefore, unless all these details are known, it is not possible to put a number on the minimum sample size.

Parady, G., K.W. Axhausen, 2024. Size matters: the use and misuse of statistical significance in discrete choice models in the transportation academic. Transportation (2024) 51:2393–2425 <https://doi.org/10.1007/s11116-023-10423-y>

Statistical power is a function of sample size, statistical significance and more importantly, effect size. That is, given a statistical significance level α , the smaller the effect, the larger the sample required to detect it. Its most common uses are to evaluate the power a statistical test had on a completed study, and to calculate necessary sample sizes given anticipated effect sizes and power (Cohen 1988). In other words, it is used to answer two questions: (a) assuming that the effect we are looking for actually exists and has magnitude m , for sample size n , what is the probability we will detect such effect (i.e., correctly reject the null) at significance level α ? And (b) what sample size do we need in our study to identify an effect of magnitude m , at significance level α with power level $1 - \beta$?

Finally, regarding sample size determination, it must be pointed out that while there is a comprehensive literature dealing with sample size for discrete choice experiments (Rose et al. 2008; Rose and Bliemer 2013) existing theory largely ignores the issue of minimum sample size requirements in terms of power (de Bekker-Grob et al. 2015).

Rose, J.M., M.C.J. Bliemer, D.A. Hensher, A.T. Collins, 2008. Designing efficient stated choice experiments in the presence of reference alternatives. Transportation Research Part B 42 (2008) 395–406. doi:10.1016/j.trb.2007.09.002

Within Eq. (8), the presence of M suggests that the (co)variances become smaller with larger sample sizes. This also follows for the asymptotic standard errors, obtained by taking the square root of the diagonal elements (including M) of this matrix (i.e., variances). By taking the square root of M , one will observe diminishing improvements to the standard errors over increases in sample size. The AVC matrix plays an important role when determining efficient experimental designs, as will be shown in the next section.

Sriwastava, A., P. Reichert, 2023. Reducing sample size requirements by extending discrete choice experiments to indifference elicitation. Journal of Choice Modelling 48 (2023) 100426. <https://doi.org/10.1016/j.jocm.2023.100426>

The small amount of information extracted from each reply is usually compensated by large sample sizes used in the “experiment” which can often relatively easily be achieved, in particular for the elicitation of societal preferences for which the number of stakeholders is beyond critical limits. However, if the problem to be assessed requires the elicitation of preferences of experts, e.g. in environmental management or health economics, it can be difficult to get a sufficiently large sample size for obtaining reliable preference estimates (de Bekker-Grob et al., 2015).

Website: <https://medium.com/@gowthamiarva28/from-zero-to-hero-logistic-regression-made-it-simple-3eb1825cf1e9>

Jahan, M.D., T. Bhowmik, L. Hoover, N. Eluru, 2025. Comparing the Performance of Different Missing Data Imputation Approaches in Discrete Outcome Modeling. Transportation Research Record, Vol. 2679(2), 879–903.

The simulation results across larger sample sizes are found to be consistent. Therefore, to conserve space, the results of the 500, 1,000, and 2,000-observation samples are presented. The readers should recognize that estimating discrete outcome models implicitly assumes that model

parameters converge asymptotically. However, depending on the characteristics of the dependent and independent variables, the sample size requirements for asymptotic convergence and parameter stability could vary substantially. Interested readers can explore earlier works on sample size requirements for their specific dataset following guidelines from earlier research (39–42).

Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, A.R. Feinstein, 1996. A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. *Clinical Epidemiology*, Vol. 49, No. 12, 1996, pp. 1373–1379. <https://doi.org/10.1016/j.amepre.2003.12.002>

We performed a Monte Carlo study to evaluate the effect of the number of events per variable (EPV) analyzed in logistic regression analysis. The simulations were based on data from a cardiac trial of 673 patients in which 252 deaths occurred and seven variables were cogent predictors of mortality; the number of events per predictive variable was $(252/7=)$ 36 for the full sample. For the simulations, at values of $EPV = 2, 5, 10, 15, 20,$ and $25,$ we randomly generated 500 samples of the 673 patients, chosen with replacement, according to a logistic model derived from the full sample. Simulation results for the regression coefficients for each variable in each group of 500 samples were compared for bias, precision, and significance testing against the results of the model fitted to the original sample. For EPV values of 10 or greater, no major problems occurred. For EPV values less than 10, however, the regression coefficients were biased in both positive and negative directions; the large sample variance estimates from the logistic model both overestimated and underestimated the sample variance of the regression coefficients; the 90% confidence limits about the estimated values did not have proper coverage; the Wald statistic was conservative under the null hypothesis; and paradoxical associations (significance in the wrong direction) were increased. Although other factors (such as the total number of events, or sample size) may influence the validity of the logistic model, our findings indicate that low EPV can lead to major problems.

Riley, R.D., K.I.E. Snell, J. Ensor, D.L. Burke, F.E. Harrell Jr, K.G.M. Moons, G.S. Collins, 2019. Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. *Statistics in Medicine*. Vol. 38, No.7, 1262-127. <https://doi.org/10.1002/sim.7993>

This paper is for continuous distribution, but it is tied into their Part II paper above.

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 the following 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 prespecification 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 criteria 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.

Van Calster, B., M. van Smeden, B. De Cock, E.W. Steyerberg, 2020. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. *Statistical Methods in Medical Research*, Vol. 29, No. 11. 3166–3178. DOI: 10.1177/0962280220921415

When developing risk prediction models on datasets with limited sample size, shrinkage methods are recommended. Earlier studies showed that shrinkage results in better predictive performance on average. This simulation study aimed to investigate the variability of regression shrinkage on predictive performance for a binary outcome. We compared standard maximum likelihood with the following shrinkage methods: uniform shrinkage (likelihood-based and bootstrap-based), penalized maximum likelihood (ridge) methods, LASSO logistic regression, adaptive LASSO, and Firth's correction. In the simulation study, we varied the number of predictors and their strength, the correlation between predictors, the event rate of the outcome, and the events per variable. In terms of results, we focused on the calibration slope. The slope indicates whether risk predictions are too extreme (slope < 1) or not extreme enough (slope > 1). The results can be summarized into three main findings. First, shrinkage improved calibration slopes on average. Second, the between sample variability of calibration slopes was often increased relative to maximum likelihood. In contrast to other shrinkage approaches, Firth's correction had a small shrinkage effect but showed low variability. Third, the correlation between the estimated shrinkage and the optimal shrinkage to remove overfitting was typically negative, with Firth's correction as the exception. We conclude that, despite improved performance on average, shrinkage often worked poorly in individual datasets, in particular when it was most needed. The results imply that shrinkage methods do not solve problems associated with small sample size or low number of events per variable.

Pavlou, M., G. Ambler, C. Qu, S.R. Seaman, I.R. White, R.Z. Omar, 2024. An evaluation of sample size requirements for developing risk prediction models with binary outcomes. *BMC Medical Research Methodology* Vol. 24, 146. <https://doi.org/10.1186/s12874-024-02268-5>

In practice, it is important that the sample size be chosen with the clinical aims of the model in mind. The RvS formulae investigated in this paper are important because they consider two important aspects of predictive performance: calibration and predictive accuracy. However, they only target average values of calibration slope and MAPE and there is, of course, no guarantee that an individual model fitted on an adequately sized sample from the target population will achieve these values. Even in cases where a calibration target is met on average, the variability in the calibration slope can be quite high. One such scenario we have seen in this article is when the number of candidate predictor variables is smaller than 10. Our simulation-based approach, implemented in the software 'samplesizedev', in addition to estimating the sample size required to achieve a target calibration slope on average, also allows quantification of the variability in the calibration slope for that sample size.

Dhiman, P., J. Ma, C. Qi, G. Bullock, J.C. Sergeant, R.D. Riley, G.S. Collins, 2023. Sample size requirements are not being considered in studies developing prediction models for binary outcomes: a systematic review. *BMC Medical Research Methodology* (2023) 23:188. <https://doi.org/10.1186/s12874-023-02008-1>

Background

Having an appropriate sample size is important when developing a clinical prediction model. We aimed to review how sample size is considered in studies developing a prediction model for a binary outcome.

Methods

We searched PubMed for studies published between 01/07/2020 and 30/07/2020 and reviewed the sample size calculations used to develop the prediction models. Using the available information, we calculated the minimum sample size that would be needed to estimate overall risk and minimise overfitting in each study and summarised the difference between the calculated and used sample size.

Results

A total of 119 studies were included, of which nine studies provided sample size justification (8%). The recommended minimum sample size could be calculated for 94 studies: 73% (95% CI: 63–82%) used sample sizes lower than required to estimate overall risk and minimise overfitting including 26% studies that used sample sizes lower than required to estimate overall risk only. A similar number of studies did not meet the ≥ 10 EPV criteria (75%, 95% CI: 66–84%). The median deficit of the number of events used to develop a model was 75 [IQR: 234 lower to 7 higher] which reduced to 63 if the total available data (before any data splitting) was used [IQR:225 lower to 7 higher]. Studies that met the minimum required sample size had a median c-statistic of 0.84 (IQR:0.80 to 0.9) and studies where the minimum sample size was not met had a median c-statistic of 0.83 (IQR: 0.75 to 0.9). Studies that met the ≥ 10 EPP criteria had a median c-statistic of 0.80 (IQR: 0.73 to 0.84).

Conclusions

Prediction models are often developed with no sample size calculation, as a consequence many are too small to precisely estimate the overall risk. We encourage researchers to justify, perform and report sample size calculations when developing a prediction model.

Hossain, M.B., M. Sadatsafavi, J.C. Johnston, H. Wong, V.J. Cook, M.E. Karim, 2025. LASSO-Based Survival Prediction Modeling with Multiply Imputed Data: A Case Study in Tuberculosis Mortality Prediction. The American Statistician. Vol. 00, No. 0, 1–12: Statistical Practice. <https://doi.org/10.1080/00031305.2025.252654>

We used the sample size determination technique for the primary analysis with a survival outcome (Riley et al. 2019). With a sample size of 2923 TB survivors, our model could incorporate 34 parameters (Appendix 1D). However, we had 44 parameters (36 main effects and 8 interactions) for our model. To deal with overfitting and internally validate the model, we applied Cox-LASSO shrinkage with 5-fold cross-validation. For each fold, the lambda hyperparameter was chosen using 5-fold cross-validation. We considered two techniques to choose the hyperparameter: lambda that gave the minimum cross-validated prediction error (“minimum-lambda”), and lambda within one standard error (SE) from the minimum (“1SE-lambda”). The minimum-lambda is the commonly used approach (Simon et al. 2011), while the latter approach could produce a more regularized and parsimonious model (Krstajic et al. 2014).

Sample Size Requirements:

Logistic regression requires a relatively large sample size to produce reliable and stable estimates. Small sample sizes can lead to overfitting and unreliable predictions.

Riley, R.D., K. I.E. Snell, G.P. Martin, R. Whittle, L. Archer, M. Sperrin, G.S. Collins, 2021b. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. Journal of Clinical Epidemiology, Vol. 132, 88-96. (<https://doi.org/10.1016/j.jclinepi.2020.12.005>)

“Objectives: When developing a clinical prediction model, penalization techniques are recommended to address overfitting, as they shrink predictor effect estimates toward the null and reduce mean-square prediction error in new individuals. However, shrinkage and penalty terms (‘tuning parameters’) are estimated with uncertainty from the development data set. We examined the magnitude of this uncertainty and the subsequent impact on prediction model performance.

Study Design and Setting: This study comprises applied examples and a simulation study of the following methods: uniform shrinkage (estimated via a closed-form solution or bootstrapping), ridge regression, the lasso, and elastic net.

Results: In a particular model development data set, penalization methods can be unreliable because tuning parameters are estimated with large uncertainty. This is of most concern when development data sets have a small effective sample size and the model’s Cox-Snell R^2 is low. The problem can lead to considerable miscalibration of model predictions in new individuals.

Conclusion: Penalization methods are not a ‘carte blanche’; they do not guarantee a reliable prediction model is developed. They are more unreliable when needed most (i.e., when overfitting may be large). We recommend they are best applied with large effective sample sizes, as identified from recent sample size calculations that aim to minimize the potential for model overfitting and precisely estimate key parameters.”

Baeza-Delgado, C. L.C. Alberich, J.M. Carot-Sierra, D. Veiga-Canuto, B. M. de las Heras, B. Raza, L. Martí-Bonmatí, 2022. A practical solution to estimate the sample size required for clinical prediction models generated from observational research on data. European Radiology Experimental, Vol. 6, No. 22. (<https://doi.org/10.1186/s41747-022-00276-y>)

Background: Estimating the required sample size is crucial when developing and validating clinical prediction models. However, there is no consensus about how to determine the sample size in such a setting. Here, the goal was to compare available methods to define a practical solution to sample size estimation for clinical predictive models, as applied to Horizon 2020 PRIMAGE as a case study.

Methods: Three different methods (Riley’s; “rule of thumb” with 10 and 5 events per predictor) were employed to calculate the sample size required to develop predictive models to analyse the variation in sample size as a function of different parameters. Subsequently, the sample size for model validation was also estimated.

Results: To develop reliable predictive models, 1397 neuroblastoma patients are required, 1060 high-risk neuroblastoma patients and 1345 diffuse intrinsic pontine glioma (DIPG) patients. This

sample size can be lowered by reducing the number of variables included in the model, by including direct measures of the outcome to be predicted and/or by increasing the follow-up period. For model validation, the estimated sample size resulted to be 326 patients for neuroblastoma, 246 for high-risk neuroblastoma, and 592 for DIPG.

Conclusions: Given the variability of the different sample sizes obtained, we recommend using methods based on epidemiological data and the nature of the results, as the results are tailored to the specific clinical problem. In addition, sample size can be reduced by lowering the number of parameter predictors, by including direct measures of the outcome of interest.

Tian, Y., H. Rusinek, A.V. Masurkar, Y. Feng, 2024. ℓ_1 -Penalized Multinomial Regression: Estimation, Inference, and Prediction, With an Application to Risk Factor Identification for Different Dementia Subtypes. *Statistics in Medicine*, 2024; 43:5711–5747.
<https://doi.org/10.1002/sim.10263>

High-dimensional multinomial regression models are very useful in practice but have received less research attention than logistic regression models, especially from the perspective of statistical inference. In this work, we analyze the estimation and prediction error of the contrast-based ℓ_1 -penalized multinomial regression model and extend the debiasing method to the multinomial case, providing a valid confidence interval for each coefficient and p value of the individual hypothesis test. We also examine cases of model misspecification and non-identically distributed data to demonstrate the robustness of our method when some assumptions are violated. We apply the debiasing method to identify important predictors in the progression into dementia of different subtypes. Results from extensive simulations show the superiority of the debiasing method compared to other inference methods.

Collins, G.S., P. Dhiman, J. Ma, M.M. Schlüssel, L. Archer, B. Van Calster, F.E. Harrell Jr, G.P. Martin, K.G.M Moons, M. van Smeden, M. Sperrin, G.S. Bullock, R.D. Riley, 2024. Evaluation of clinical prediction models (part 1): from development to external validation. *British Medical Journal*, Vol. 384, e074821. <http://dx.doi.org/10.1136/>

Clinical prediction models use a combination of variables to estimate outcome risk for individuals. Evaluating the performance of a prediction model is critically important and validation studies are essential, as a poorly developed model could be harmful or exacerbate disparities in either provision of health care or subsequent healthcare outcomes.

Evaluating model performance should be carried out in datasets that are representative of the intended target populations for the model's Implementation.

A model's predictive performance will often appear to be excellent in the development dataset but be much lower when evaluated in a separate dataset, even from the same population. Splitting data at the moment of model development should generally be avoided as it discards data leading to a more unreliable model, whilst leaving too few data to reliably evaluate its performance

Concerted efforts should be made to exploit all available data to build the best possible model, with better use of resampling methods for internal validation, and internal-external validation to evaluate model performance and generalisability across clusters.

Riley, R.D., L. Archer, K.I.E. Snell, J. Ensor, P. Dhiman, G.P. Martin, L.J. Bonnett, G.S. Collins, 2024a. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. British Medical Journal, Vol. 384, e074820 <http://dx.doi.org/10.1136/>

External validation is the evaluation of a model's predictive performance in a different (but relevant) dataset, which was not used in the development process. An external validation study involves five key steps: obtaining a suitable dataset, making outcome predictions, evaluating predictive performance, assessing clinical usefulness, and clearly reporting findings.

The validation dataset should represent the target population and setting in which the model is planned to be implemented.

At a minimum, the validation dataset must contain the information needed to apply the model (ie, to make predictions) and make comparisons to observed outcomes.

A model's predictive performance should be examined in terms of overall fit, calibration, and discrimination, in the overall population and ideally in key subgroups (eg, defined by ethnic group), as part of fairness checks.

Calibration should be examined across the entire range of predicted values, and at each relevant time point for which predictions are being made, using a calibration plot including a smoothed flexible calibration curve.

Where the goal is for predictions to direct decision making, a prediction model should also be evaluated for its clinical usefulness, for example, using net benefit and decision curves. Although a well calibrated model is ideal, a miscalibrated model might still have clinical usefulness. The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement provides guidance on how to report external validation studies.

Riley, R.D., K.I.E. Snell, L. Archer, J. Ensor, T.P.A. Debray, B. Van Calster, M. van Smeden, G.S. Collins, 2024b. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. British Medical Journal, Vol. 384, e074821. doi: <http://dx.doi.org/10.1136/bmj-2023-074821>

The sample size for an external validation study should be large enough to precisely estimate the predictive performance of the model of interest. Many existing validation studies are too small, which leads to wide confidence intervals of performance estimates and potentially misleading claims about a model's reliability or its performance compared with other models.

To deal with concerns of imprecise performance estimates, rules of thumb for sample size have been proposed, such as having at least 100 events and 100 non-events.

Such rules of thumb provide a starting point but are problematic, because they are not specific to either the model or the clinical setting, and precision also depends on factors other than the number of events and non-events. A more tailored approach can allow researchers to calculate the sample size required to target chosen precision (confidence interval widths) of key performance estimates, such as for R², calibration curve, c statistic, and net benefit.

Calculations depend on users specifying information such as the outcome proportion, expected model performance, and distribution of predicted values, which can be gauged from the original model development study. The `pmvalsampsiz` package in Stata and R allows researchers to implement the approach with one line of code.

Heckmann, T., K. Gegg, A. Gegg, M. Becht, 2014. Sample size matters: investigating the effect of sample size on a logistic regression susceptibility model for debris flows. *Natural Hazards and Earth System Sciences*, Vol. 14, 259–278. doi:10.5194/nhess-14-259-2014

Abstract. Predictive spatial modelling is an important task in natural hazard assessment and regionalisation of geomorphic processes or landforms. Logistic regression is a multivariate statistical approach frequently used in predictive modelling; it can be conducted stepwise in order to select from a number of candidate independent variables those that lead to the best model. In our case study on a debris flow susceptibility model, we investigate the sensitivity of model selection and quality to different sample sizes in light of the following problem: on the one hand, a sample has to be large enough to cover the variability of geofactors within the study area, and to yield stable and reproducible results; on the other hand, the sample must not be too large, because a large sample is likely to violate the assumption of independent observations due to spatial autocorrelation. Using stepwise model selection with 1000 random samples for a number of sample sizes between $n = 50$ and $n = 5000$, we investigate the inclusion and exclusion of geofactors and the diversity of the resulting models as a function of sample size; the multiplicity of different models is assessed using numerical indices borrowed from information theory and biodiversity research. Model diversity decreases with increasing sample size and reaches either a local minimum or a plateau; even larger sample sizes do not further reduce it, and they approach the upper limit of sample size given, in this study, by the autocorrelation range of the spatial data sets. In this way, an optimised sample size can be derived from an exploratory analysis. Model uncertainty due to sampling and model selection, and its predictive ability, are explored statistically and spatially through the example of 100 models estimated in one study area and validated in a neighbouring area: depending on the study area and on sample size, the predicted probabilities for debris flow release differed, on average, by 7 to 23 percentage points. In view of these results, we argue that researchers applying model selection should explore the behaviour of the model selection for different sample sizes, and that consensus models created from a number of random samples should be given preference over models relying on a single sample.

Nemes, S., J.M. Jonasson, A. Genell, G. Steineck, 2009. Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology* 2009, 9:56 doi:10.1186/1471-2288-9-56

Abstract

Background: In epidemiological studies researchers use logistic regression as an analytical tool to study the association of a binary outcome to a set of possible exposures.

Methods: Using a simulation study we illustrate how the analytically derived bias of odds ratios modelling in logistic regression varies as a function of the sample size.

Results: Logistic regression overestimates odds ratios in studies with small to moderate samples size. The small sample size induced bias is a systematic one, bias away from null. Regression coefficient estimates shifts away from zero, odds ratios from one.

Conclusion: If several small studies are pooled without consideration of the bias introduced by the inherent mathematical properties of the logistic regression model, researchers may be misled to erroneous interpretation of the results.

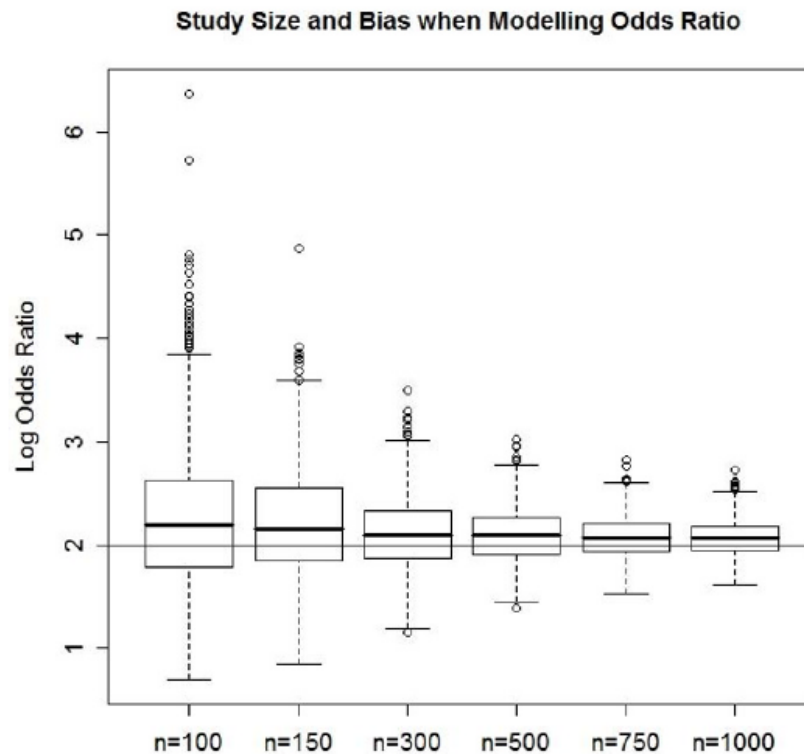


Figure 3
Increasing sample size not only reduces the analytically induced bias in regression estimates but protects against extreme value estimates.

Tervonen, T., F. Pignatti, D. Postmus, 2019. From Individual to Population Preferences: Comparison of Discrete Choice and Dirichlet Models for Treatment Benefit-Risk Tradeoffs. Medical Decision Making, Vol. 39 (7), 879–885. DOI: 10.1177/0272989X19873630

Results. Population preference estimates of all models were very close to the sample mean, and the MNL and MXL models had good fit (McFadden's adjusted $R^2 = 0.12$ and 0.13). The Dirichlet model converged reliably to within 0.05 distance of the population preference estimates with a sample size of 100, where the MNL model required a sample size of 240 for this (note: 3 variables). The MNL model produced consistently significant coefficient estimates with sample sizes of 100 and higher. Conclusion. The Dirichlet model is likely to have smaller sample size requirements than standard discrete choice models in modeling population preferences for treatment benefit-risk tradeoffs and is a useful addition to health preference analyst's toolbox.

Edlinger, M., M. van Smeden, H.F. Alber, M. Wanitschek, B. Van Calster, 2022. Risk prediction models for discrete ordinal outcomes: Calibration and the impact of the proportional odds assumption. *Statistics in Medicine*. Vol. 41, 1334–1360. DOI: 10.1002/sim.9281

Using large sample simulations, we studied the performance of models for risk estimation under various conditions, assuming that the true model has either a multinomial logistic form or a cumulative logit proportional odds form. Small sample simulations were used to compare the tendency for overfitting between models. As a case study, we developed models to diagnose the degree of coronary artery disease (five categories) in symptomatic patients. When the true model was multinomial logistic, proportional odds models often yielded poor risk estimates, with calibration slopes deviating considerably from unity even on large model development datasets. The stereotype logistic model improved the calibration slope, but still provided biased risk estimates for individual patients. When the true model had a cumulative logit proportional odds form, multinomial logistic regression provided biased risk estimates, although these biases were modest. Nonproportional odds models require more parameters to be estimated from the data, and hence suffered more from overfitting. Despite larger sample size requirements, we generally recommend multinomial logistic regression for risk prediction modeling of discrete ordinal outcomes.

Diomatari, C., G.P. Martin, D. A. Jenkins, M. Jani, 2025. Clinical prediction models for medication adverse events in patients with rheumatic and musculoskeletal conditions: A systematic literature review. *Seminars in Arthritis and Rheumatism*, Vol. 73, 152728.

Results: Of 2406 studies identified, 1734 titles/abstracts were screened, and 38 were reviewed in full. Twelve studies reporting 17 CPMs met eligibility criteria. Most CPMs (76.4 %) focused on rheumatoid arthritis and disease modifying anti-rheumatic drugs (DMARDs) such as methotrexate (69.2 %) and biologic drugs (15.3 %). Cox proportional hazards or logistic regression models were commonly used. Twelve models (70.5 %) had high overall ROB due to inappropriate variable selection methods and sample size.

Conclusions: This is the first systematic review summarising CPMs for AEs associated with RMD medications. It highlights that existing CPMs are affected by methodological pitfalls, including inappropriate variable selection and lack of clear sample size justification. Future models could consider a broader range of RMDs and medications. Emerging methods such as machine learning with the ability to model complex interactions, and multioutcome CPMs to predict several AEs to one class of drug may improve predictions.

...Developing a model with an insufficient sample size can lead to overfitting, resulting to unreliable predictions, as the model may capture noise rather than the underlying signal in the data [6]. Traditionally, rules-of-thumb based on the events per variable (EPV) have been used as justification [33]. However, this standard may not be universally sufficient, as different studies may require varying EPV values [34]. To address this, new methods have been developed to calculate the minimum required sample size based on the specific needs of individual studies, leading to more reliable findings [35]. Recognizing the importance of sample size reporting and calculation, the Transparent Reporting of a multivariable model for Individual Prognosis Or Diagnosis (TRIPOD) 2015 statement also emphasizes its inclusion in model development [36],...

Falke, A., H. Hruschka, 2017. Setting prices in mixed logit model designs. *Marketing letters*, Vol. 28, 139–154: DOI 10.1007/s11002-015-9396-4

Abstract We investigate different procedures to set prices in designs for choice based conjoint analysis using the mixed logit model which captures latent consumer heterogeneity. Besides discrete attributes, we include a linear price term in the deterministic utility function thereby treating price as continuous variable. We consider two different price intervals and several price sets which contain either two or three prices. We compare these alternatives to set prices by simulating choices for different constellations on the basis of the mixed logit model. Furthermore, we generate ten designs simultaneously instead of just one. Using these simulated choices, we estimate the parameters of the mixed logit model in the next step. To reduce the needed sample size and computation time caused by accounting for latent consumer heterogeneity, we apply Halton draws and set a minimum potential design for prior draws. ANOVA with root mean squared error between estimated and true price coefficient values of individual consumers as dependent variable shows that using more extreme prices as interval bounds and one intermediate price positioned to the right of the interval performs best.

Riley, R.D., J. Ensor, K.I.E. Snell, L. Archer, R. Whittle, P. Dhiman, J. Alderman, X. Liu, L. Kirton, J. Manson-Whitton, M. van Smeden, K.G. Moons, K. Nirantharakumar, J.-B. Cazier, A.K. Denniston, B. Van Calster, G.S. Collins, 2025. Importance of sample size on the quality and utility of AI-based prediction models for healthcare. *Lancet Digit Health*, Vol. 7, 100857. <https://doi.org/10.1016/j.landig.2025.01.013>

Rigorous study design and analytical standards are required to generate reliable findings in healthcare from artificial intelligence (AI) research. One crucial but often overlooked aspect is the determination of appropriate sample sizes for studies developing AI-based prediction models for individual diagnosis or prognosis. Specifically, the number of participants and outcome events required in datasets for model training and evaluation remains inadequately addressed. Most AI studies do not provide a rationale for their chosen sample sizes and frequently rely on datasets that are inadequate for training or evaluating a clinical prediction model. Among the ten principles of Good Machine Learning Practice established by the US Food and Drug Administration, the UK Medicines and Healthcare products Regulatory Agency, and Health Canada, guidance on sample size is directly relevant to at least three principles. To reinforce this recommendation, we outline seven reasons why inadequate sample size negatively affects model training, evaluation, and performance. Using a range of examples, we illustrate these issues and discuss the potentially harmful consequences for patient care and clinical adoption. Additionally, we address challenges associated with increasing sample sizes in AI research and highlight existing approaches and software for calculating the minimum sample sizes required for model training and evaluation.

...Small training samples produce instability in selected model predictors,—ie, different training samples of equal size result in different selected predictors and substantial changes in how a particular predictor impacts predictions. Attempts to meaningfully explain such unstable models, therefore, become futile,¹⁹ as parameter estimates (eg, intercept and predictor effect estimates), predictor selection strategies (eg, lasso, recursive feature elimination), and post-hoc explanation methods (eg, Locally Interpretable Model-agnostic Explanations [LIME] and Shapley values [SHAP]) likewise become unstable and potentially misleading.^{20–23} Compared with regression approaches, instability tends to be greater for other AI-based methods, as these usually allow greater complexity by default and thus require larger training sample sizes.^{24,25}...

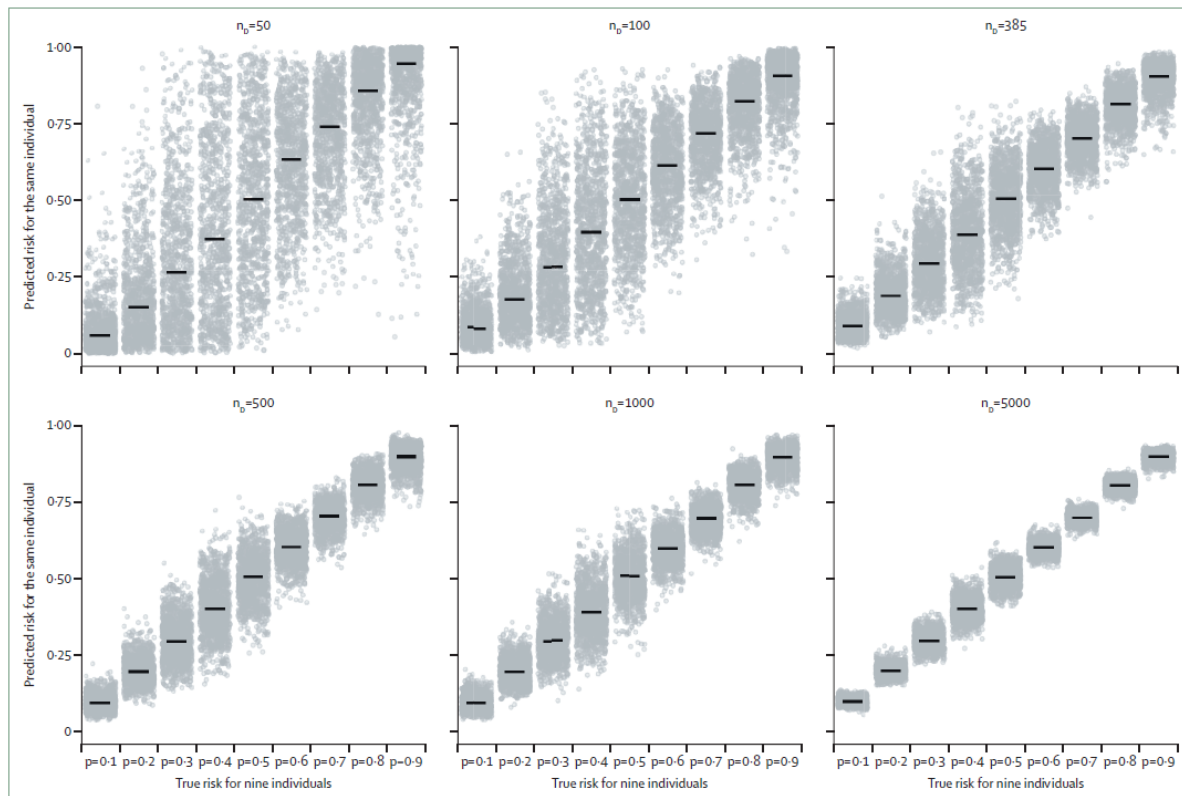


Figure 1: Instability of estimated risks across 1000 prediction models for nine individuals with true (large-sample) risks ranging from 0.1 to 0.9, based on training sample sizes (n_0) of 50, 100, 385, 500, 1000, and 5000 participants
 Each model was developed using logistic regression with a lasso penalty, fitted to a different random sample drawn from a population with an overall risk of 0.5, using one genuine predictor ($X \sim N(0,4)$) and 10 noise variables ($Z_1, \dots, Z_{10} \sim N(0,1)$). Reproduced from Riley and Collins,²⁶ with permission under a CC-BY license.

Whittle, R., J. Ensor, L. Archer, G.S. Collins, P. Dhiman, A. Denniston, J. Alderman, A. Legha, M. van Smeden, K.G. Moons, J.-B. Cazier, R.D. Riley, K.I.E. Snell, 2025. Extended sample size calculations for evaluation of prediction models using a threshold for classification. *BMC Medical Research Methodology*, Vol. 25, 170. <https://doi.org/10.1186/s12874-025-02592-4>

Abstract

When evaluating the performance of a model for individualised risk prediction, the sample size needs to be large enough to precisely estimate the performance measures of interest. Current sample size guidance is based on precisely estimating calibration, discrimination, and net benefit, which should be the first stage of calculating the minimum required sample size. However, when a clinically important threshold is used for classification, other performance measures are also often reported. We extend the previously published guidance to precisely estimate threshold-based performance measures. We have reported closed-form solutions to estimate the sample size required to target sufficiently precise estimates of accuracy, specificity, sensitivity, positive predictive value (PPV), negative predictive value (NPV), and an iterative method to estimate the sample size required to target a sufficiently precise estimate of the F1-score, in an external evaluation study of a prediction model with a binary outcome. This approach requires the user to pre-specify the target standard error and the expected value for each performance measure

alongside the outcome prevalence. We describe how the sample size formulae were derived and demonstrate their use in an example. Extension to time-to-event outcomes is also considered. In our examples, the minimum sample size required was lower than that required to precisely estimate the calibration slope, and we expect this would most often be the case. Our formulae, along with corresponding Python code and updated R, Stata and Python commands (pmvalsampsize), enable researchers to calculate the minimum sample size needed to precisely estimate threshold based performance measures in an external evaluation study. These criteria should be used alongside previously published criteria to precisely estimate the calibration, discrimination, and net-benefit.

...In each of our examples, the minimum sample size required was lower than that required to precisely estimate the calibration slope, and we expect this would most often be the case. Therefore, the precise estimation of the calibration slope should remain the focus of any sample size calculation for evaluation of a prediction model, regardless of whether a threshold is being used for classification or of the methods used to develop the model, and the formulae provided in this article should be seen as complementary to the previously published criteria [2, 5, 7]...

Legha, A., J. Ensor, R. Whittle, L. Archer, B. Van Calster, E. Christodoulou, K.I.E. Snell, M. Sadatsafavi, G.S. Collins, R.D. Riley, 2026, Sequential sample size calculations and learning curves safeguard the robust development of a clinical prediction model for individuals. Journal of Clinical Epidemiology, Vol. 191, 112117.
<https://doi.org/10.1016/j.jclinepi.2025.112117>

Abstract

Background and Objectives: When recruiting participants to a new study developing a clinical prediction model (CPM), sample size calculations are typically conducted before data collection based on sensible assumptions. This leads to a fixed sample size, but if the assumptions are inaccurate, the actual sample size required to develop a reliable model may be higher or even lower. To safeguard against this, adaptive sample size approaches have been proposed, based on sequential evaluation of (changes in) a model's predictive performance. The objective of the study was to illustrate and extend sequential sample size calculations for CPM development by (i) proposing stopping rules for prospective data collection based on minimizing uncertainty (instability) and misclassification of individual-level predictions and (ii) showcasing how it safeguards against inaccurate fixed sample size calculations.

Methods: Using the sequential approach repeats the predefined model development strategy every time a chosen number (eg, 100) of participants are recruited and adequately followed up. At each stage, CPM performance is evaluated using bootstrapping, leading to prediction and classification stability statistics and plots, alongside optimism-adjusted measures of calibration and discrimination. Learning curves display the trend of results against sample size and recruitment is stopped when a chosen stopping rule is met.

Results: Our approach is illustrated for model development of acute kidney injury using (penalized) logistic regression CPMs. Before recruitment based on perceived sensible assumptions, the fixed sample size calculation suggests recruiting 342 patients to minimize overfitting; however, during data collection, the sequential approach reveals that a much larger sample size of 1100 is required

to minimize overfitting (targeting a bootstrap-corrected calibration slope ≥ 0.9). If the stopping rule criteria also target small uncertainty and misclassification probability of individual predictions, the sequential approach suggests an even larger sample size of about 1800

Conclusion: For CPM development studies involving prospective data collection, a sequential sample size approach allows users to dynamically monitor individual-level prediction and classification instability. This helps determine when enough participants have been recruited and safeguards against using inaccurate assumptions in a sample size calculation before data collection. Engagement with patients and other stakeholders is crucial to identify sensible context-specific stopping rules for robust individual predictions.

Wang, Y., A.E. Boyd, L. Rountree, Y. Ren, K. Nyhan, R. Nagar, J. Higginbottom, M.L. Ranney, H. Parikh, B. Mukherjee, 2026. Ten Core Concepts for Ensuring Data Equity in Public Health. JAMA Health Forum. 7 (1), e256031. doi:10.1001/jamahealthforum.2025.6031.

However, it is important to recognize that data equity alone does not guarantee learning or information equity, or decision equity (Figure). How much data do we need to achieve equal predictability in 2 groups? That may not be just dictated by the sample size and representation but involve innovative prediction power calculations. Ensuring that data are equitably collected, analyzed, predicted, and interpreted is only one part of the broader ecosystem; health outcomes also depend on how information is communicated, what agency an individual or population has, and how public health decisions are made. Information and knowledge alone do not grant an individual or a population the agency to take decisions around their health.

Riley, R.D., T.P.A. Debray, G.S. Collins, L. Archer, J. Ensor, M. van Smeden, K.I.E. Snell, 2021a. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Statistics in Medicine Vol. 40, No. 19., pp. 4230-4251.

<https://doi.org/10.1002/sim.9025>

Abstract

In prediction model research, external validation is needed to examine an existing model's performance using data independent to that for model development. Current external validation studies often suffer from small sample sizes and consequently imprecise predictive performance estimates. To address this, we propose how to determine the minimum sample size needed for a new external validation study of a prediction model for a binary outcome. Our calculations aim to precisely estimate calibration (Observed/Expected and calibration slope), discrimination (C-statistic), and clinical utility (net benefit). For each measure, we propose closed-form and iterative solutions for calculating the minimum sample size required. These require specifying: (i) target SEs (confidence interval widths) for each estimate of interest, (ii) the anticipated outcome event proportion in the validation population, (iii) the prediction model's anticipated (mis) calibration and variance of linear predictor values in the validation population, and (iv) potential risk thresholds for clinical decision-making. The calculations can also be used to inform whether the sample size of an existing (already collected) dataset is adequate for external validation. We illustrate our proposal for external validation of a prediction model for mechanical heart valve failure with an expected outcome event proportion of 0.018. Calculations suggest at least 9835 participants (177 events) are required to precisely estimate the calibration and discrimination measures, with this number driven by the calibration slope criterion, which we anticipate will often be the case. Also,

6443 participants (116 events) are required to precisely estimate net benefit at a risk threshold of 8%. Software code is provided.

Sadatsafavi, M., P. Gustafson, S. Setayeshgar, L. Wynants, R.D. Riley, 2026. Bayesian Sample Size Calculations for External Validation Studies of Risk Prediction Models. Statistics in Medicine, Vol. 45, e70389. <https://doi.org/10.1002/sim.70389>

Note: this paper is the Bayesian version of the 2021 paper by Riley et al. just above. They used the same dataset.

ABSTRACT

Contemporary sample size calculations for external validation of risk prediction models require users to specify fixed values of assumed model performance metrics alongside target precision levels (e.g., 95% CI widths). However, due to the finite samples of previous studies, our knowledge of true model performance in the target population is uncertain, and so choosing fixed values represents an incomplete picture. As well, for net benefit (NB) as a measure of clinical utility, the relevance of conventional precision-based inference is doubtful. In this work, we propose a general Bayesian framework for multi-criteria sample size considerations for prediction models for binary outcomes. For statistical metrics of performance (e.g., discrimination and calibration), we propose sample size rules that target desired expected precision or desired assurance probability that the precision criteria will be satisfied. For NB, we propose rules based on Optimality Assurance (the probability that the planned study correctly identifies the optimal strategy) and Value of Information (Vol) analysis, which quantifies the expected gain in NB by learning about model performance from a validation study of a given size. We showcase these developments in a case study on the validation of a risk prediction model for deterioration among hospitalized COVID-19 patients. Compared to conventional sample size calculation methods, a Bayesian approach requires explicit quantification of uncertainty around model performance, and thereby enables flexible sample size rules based on expected precision, assurance probabilities, and Vol. In our case study, calculations based on Vol for NB suggest considerably lower sample sizes are required than when focusing on the precision of calibration metrics. This approach is implemented in the accompanying software.

Figure 2 shows the exemplary kernel histograms of the distribution of CI widths (the first three panels) for the smallest ($N = 306$) and largest ($N = 1181$) components of the sample size. The last panel demonstrates the distribution of the incremental NB of the model compared with the default strategies (i.e., $NB1 - \max(NB0 - NB2)$). With higher sample sizes, the CI widths get both shorter and more clustered. The distributions are relatively symmetrical. This can explain why the ECIW values are close to their conventional, frequentist counterparts. For NB, as the sample estimate of NB is an unbiased estimator, higher sample sizes will result in a narrower distribution, but their location remains the same.

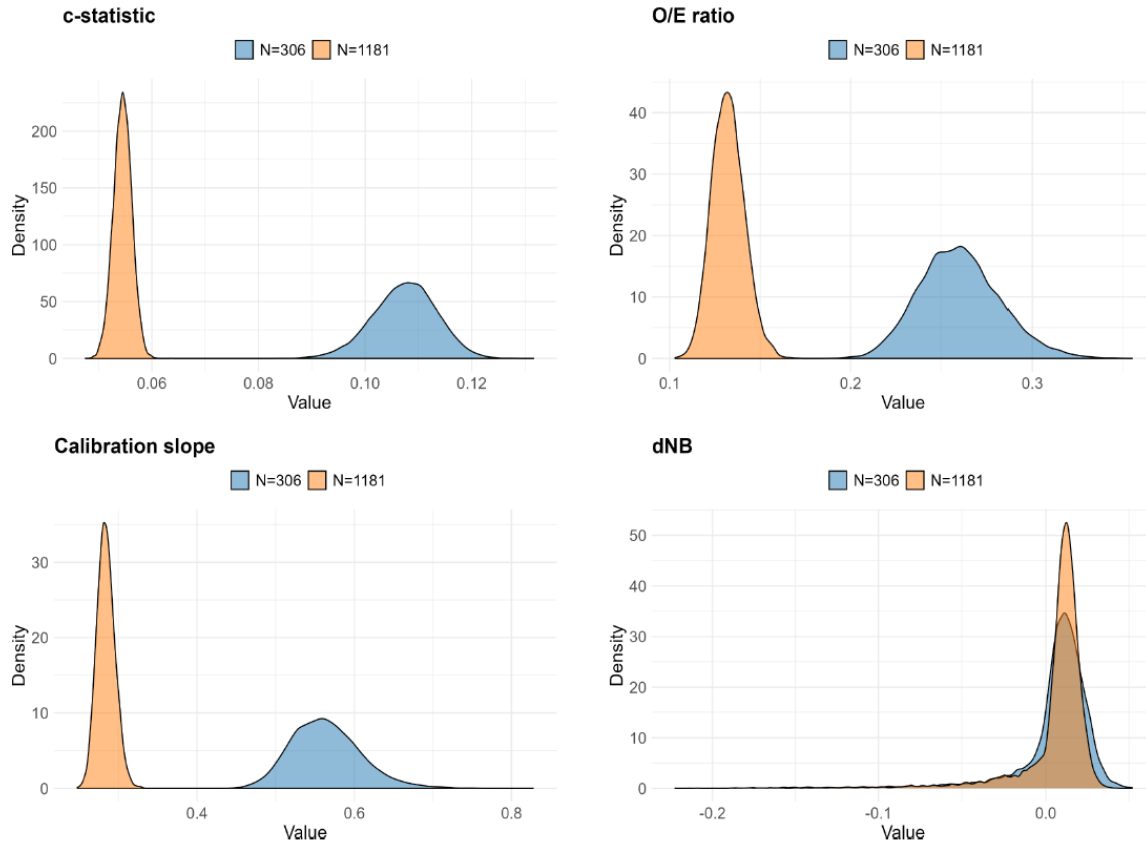


FIGURE 2 | Kernel histograms of CI widths for (a) c-statistic, (b) O/E ratio, and (c) calibration slope. Also, (d) shows the kernel histogram of the incremental net benefit of the model compared to the best default strategy for two sample sizes.

Additional References

Babayak, M.J., 2004. What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. *Psychosomatic Medicine*, Vol. 66, pp. 411–421 DOI: 0033-3174/04/6603-0411 (<https://people.duke.edu/~mababayak/papers/babayakregression.pdf>)

Bai, Z., D. Jiang, J.-F. Yao, S. Zheng, 2009. Corrections to LRT on large-dimensional covariance matrix by RMT. *The Annals of Statistics*, Vol. 37, No. 6B, 3822–3840. DOI: 10.1214/09-AOS694

Cheng, Y., K.V. Petrides, J. Li, 2025. Estimating the Minimum Sample Size for Neural Network Model Fitting—A Monte Carlo Simulation Study. *Behavioral Sciences*, Vol. 15, No. 2, 211. <https://doi.org/10.3390/bs15020211>

He, Y., T Jiang, J. Wen, G. Xu, 2020. Likelihood Ratio Test in Multivariate Linear Regression: from Low to High Dimension. *Statistica Sinica*, Vol. 46, 8479–8492.

He, Y., Z. Wang, G. Xu, 2021. A Note on the Likelihood Ratio Test in High-Dimensional Exploratory Factor Analysis. *Psychometrika*, Vol. 86, No. 2, 442–463. doi: 10.1007/s11336-021-09755-4

Hensher, D.A., A.T. Collins, W.H. Greene, 2013. Accounting for attribute non-attendance

and common-metric aggregation in a probabilistic decision process mixed multinomial logit model: a warning on potential confounding. *Transportation*, 40, pp.1003–1020. DOI 10.1007/s11116-012-9447-0.

Hosmer, D.W., S. Lemeshow, 2000. *Applied Logistic Regression*, 2nd Edition. John Wiley & Sons Inc. New York, N.Y.

Hughes, V., 2017. Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough? *Speech Communication*, Volume 94, pp. 15-29.

<https://doi.org/10.1016/j.specom.2017.08.005>

Lever, J., M. Krzywinski, N. Altman, 2016. Model selection and overfitting. *Nature Methods*, Vol. 13, No. 9, 703-704.

Melvin, R.L., 2021. Sample Size in Machine Learning and Artificial Intelligence. Blog Post (<https://sites.uab.edu/periop-datascience/2021/06/28/sample-size-in-machine-learning-and-artificial-intelligence/>)

Nibbering, D., 2024. A high-dimensional multinomial logit model. *Journal of Applied Econometrics*. No. 4, pp. 481-497. <https://doi.org/10.1002/jae.3034>

Rigg, J., M. Hankins, 2015. MO1 - Reducing And Quantifying Over-Fitting In Regression Models. *Research on Modeling Methods Studies*, Vol. 18, No. 3, pA. 10.1016/j.jval.2015.03.027

Rold, J.A., S.B. Sidaty-Regad, 2021. Comparison of the likelihood ratios of two diagnostic tests subject to a paired to a paired design: confidence interval and sample size. *REVSTAT – Statistical Journal*, Volume 19, Number 4, 575–601.

Royle, J.R., R.M. Dorazio, 2009. Chapter 2 – Essentials of Statistical Inference. In *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*, 27-82. <https://doi.org/10.1016/B978-0-12-374097-7.00004-1>

Silvey, S., A. Olex, S. Tang, J. Liu, 2026. Sample Size Requirements for Machine Learning Classification of Binary Outcomes in Bulk RNA-Seq Data. *BMC Bioinformatics*, Vol. 27, No. 53. <https://doi.org/10.1101/2025.08.19.25333999>

Thankappan, K., 2024. Sample Size in Machine Learning. Blog Post. https://medium.com/@drkrishnakumart_51860/sample-size-in-machine-learning-74d5aa6eb321

Thomassen, D., T. Hackmann, J. Goeman, E. Steyerberg, S. le Cessie, 2025. Effective sample size for individual risk predictions: quantifying uncertainty in machine learning models. *Lancet Digit Health*, Vol. 7, 100911. <https://doi.org/10.1016/j.landig.2025.100911>

Wang, S., Q. Wang, N. Bailey, J. Zhao, 2021. Deep neural networks for choice analysis: A statistical learning theory perspective. *Transportation Research Part B*, Vol. 148, 60-81.

Ye, F., D. Lord, 2014. Comparing Three Commonly Used Crash Severity Models on Sample Size Requirements: Multinomial Logit, Ordered Probit and Mixed Logit Models. *Analytic Methods in Accident Research*, Vol. 1, 72-85.

Yuan, K.H., C. Fana, Y. Zhao, 2019. What Causes the Mean Bias of the Likelihood Ratio Statistic with Many Variables? *Multivariate Behavioral Research*, Vol. 54, No. 6, 840–855.
<https://doi.org/10.1080/00273171.2019.1596060>

Zabor, E.C., C.A. Reddy, R.D. Tendulkar, S. Patil, 2022, Logistic Regression in Clinical Studies. *International Journal of Radiation Oncology Biology Physics*, Vol. 112, No. 2, pp. 271–277.

Zamagni, G. C. Fregona, M. Barbieri, M.S. Scalia, L. Monasta, C. Lees, T. Stampalija, G. Barbati, 2026. Assessing adherence to TRIPOD+AI guidelines in machine learning models for predicting small for gestational age and fetal growth restriction: a systematic review. *American Journal of Obstetrics & Gynecology MFM*, Vol. 8, 101862. <http://dx.doi.org/10.1016/j.ajogmf.2025.1018>

Zantvoort, K., B. Nacke, D. Görlich, S. Hornstein, C. Jacobi, B. Funk, 2024. Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions. *npj Digital Medicine*, Vol. 7, 361 <https://doi.org/10.1038/s41746-024-01360-w>

Zhao, X., X. Yan, A. Yu, P. Van Hentenryck, 2019. Modeling Stated Preference for Mobility-on-Demand Transit: A Comparison of Machine Learning and Logit Models.
<https://arxiv.org/pdf/1811.01315>

Acknowledgements

I would like to thank all my colleagues and friends who have provided feedback over the last few months.