A simulation approach to calculating minimum sample sizes for prediction modelling

The pmsims package for R

Ewan Carr, Gordon Forbes, Diana Shamsutdinova, Daniel Stahl, and Felix Zimmer

Department of Biostatistics & Health Informatics
King's College London

29th August 2023

30-second version

1. Prediction models developed with inadequate samples lead to overfitting and imprecise estimates.

2. Existing tools use analytical methods to derive minimum samples sizes for continuous, binary, and survival outcomes.

3. We've developed a simulation-based approach that can be applied to any outcome or method.

Overview

1. Background
    - What's the problem?
    - What solutions already exist?
2. Our approach
    - Simulation to identify minimum sample size that satisfies criteria.
    - Flexible, but slower.
    - Gaussian process regression (via `mlpwr`) to speed up.
3. Next steps

What's the problem?

- Thousands of prediction models are developed every year.
- Prediction models can inform treatment decisions, facilitate screening, and enable stratified care.
- However, most are developed with inadequate samples.

- In a review of prediction models for COVID-19, the most frequent problem was insufficient sample size. 67% (408/731) of models were developed on too few patients.[1]
- 56% of models developed using supervised machine learning techniques are developed using inadequate sample sizes.[2]
- 73% of prediction models for psychiatry had inadequate sample size.[4]
- 8% of machine learning models published in oncology report a sample size justification.[3]

Inadequate samples → research waste

- Inadequate samples lead to overfitting and inaccurate estimates of model parameters.
    - Overfitting is where the model captures idiosyncrasies of the development sample, producing inflated estimates of predictive performance that cannot be replicated in the target population.
- Unreliable models may generate inappropriate decisions about patient care or lead to models not being implemented into clinical practice.
- Data collection can be invasive and inconvenient and diverts resources from other activities that benefit patients.

Ensuring sample sizes are sufficient *before model development* would improve patient outcomes by avoiding models developed with inadequate samples and reducing participant burden.

Tools for estimating minimum sample sizes for prediction

Until recently, most studies ignored sample size.

Or they used simple rules-of-thumb (e.g., 10 events per variable).

In 2018, `pmsampsize` was released by Riley et al.

RESEARCH ARTICLE

WILEY Statistics in Medicine

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

Richard D. Riley[1] | Kym I.E. Snell[1] | Joie Ensor[1] | Danielle L. Burke[1] | Frank E. Harrell Jr[2] | Karel G.M. Moons[3] | Gary S. Collins[4]

[1] Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University, Keele, UK

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

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

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

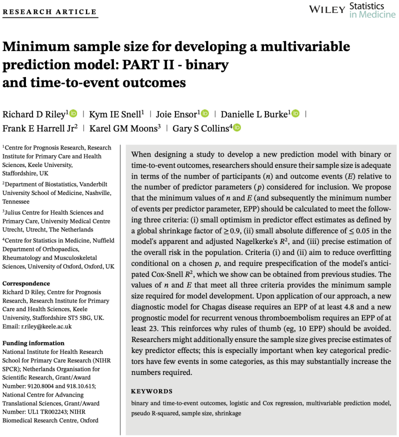
Correspondence
Richard D Riley, Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University, Keele, Staffordshire ST5 5BG, UK.
Email: r.riley@keele.ac.uk

Funding information
National Institute for Health Research School for Primary Care Research (NIHR SPCR); Netherlands Organisation for Scientific Research, Grant/Award Number: project 9120.8004 and 918.10.615; CTSA, Grant/Award Number: UL1 TR002243; National Center for Advancing Translational Sciences; US National Institutes of Health; NIHR Biomedical Research Centre

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.

KEYWORDS

continuous outcome, linear regression, minimum sample size, multivariable prediction model, R-squared



RESEARCH ARTICLE

WILEY Statistics in Medicine

Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes

Richard D Riley[1] | Kym IE Snell[1] | Joie Ensor[1] | Danielle L Burke[1] | Frank E Harrell Jr[2] | Karel GM Moons[3] | Gary S Collins[4]

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

[2] Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, Tennessee

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

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

Correspondence
Richard D Riley, Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire ST5 5BG, UK.
Email: r.riley@keele.ac.uk

Funding information
National Institute for Health Research School for Primary Care Research (NIHR SPCR); Netherlands Organisation for Scientific Research, Grant/Award Number: 9120.8004 and 918.10.615; National Centre for Advancing Translational Sciences, Grant/Award Number: UL1 TR002243; NIHR Biomedical Research Centre, Oxford

When designing a study to develop a new prediction model with binary or time-to-event outcomes, researchers should ensure their sample size is adequate in terms of the number of participants (n) and outcome events (E) relative to the number of predictor parameters (p) considered for inclusion. 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 (eg, 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.

KEYWORDS

binary and time-to-event outcomes, logistic and Cox regression, multivariable prediction model, pseudo R-squared, sample size, shrinkage

pmsampsize

The package identifies the minimum sample that results in:

	Continuous	Binary
i.	Small optimism in predictor effect estimates, indicated 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 of the model’s residual standard deviation.	Precise estimation of the overall risk in the population.
iv.	Precise estimation of the model intercept.	

We ❤️ pmsampsize, however…

pmsampsize has methods for simple continuous, binary, and survival outcome. However, we increasingly need to derive minimum samples for:

Other types of model

e.g., machine learning algorithms such as random forests or gradient boosting.

Other data types

e.g., repeated measures and longitudinal data.

So, we've created a simulation-based framework for sample size estimation for prediction.

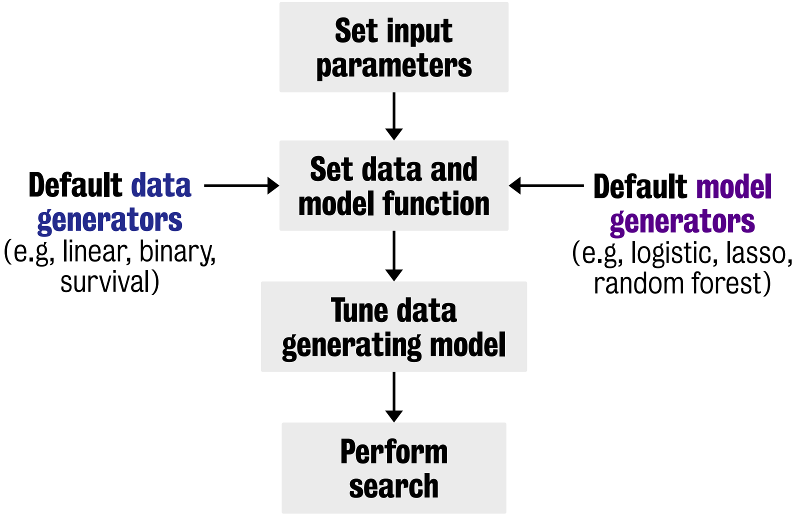
pmsims

A simulation-based framework that derives the minimum sample that

- achieves with 10% of the expected large-sample performance
- achieves calibration slope of >0.9

- Any model or data type
- Defaults for common model/data types
- Fast(er).

Our approach



Set input parameters

Set data and model function

Default data generators (e.g, linear, binary, survival)

Default model generators (e.g, logistic, lasso, random forest)

Tune data generating model

Perform search

Slide explaining input parameters

- `simulate_continuous`
- `simulate_binary`
- `simulate_survival`

Which each call:

- `simulate_custom`

Slide explaining default data and model generators

List default data/model/metrics.

Performing the search: `mlpwr`

A simulation-based approach with complex data or models would be too slow.

`mlpwr` is a R package by Felix Zimmer and Rudolf Debelak at the University of Zurich.

> *“A Power Analysis Toolbox to Find Cost-Efficient Study Designs”*

Sample Size Planning for Complex Study Designs: A Tutorial for the `mlpwr` Package

Felix Zimmer, Mirka Henninger, and Rudolf Debelak
University of Zurich

A common challenge in designing empirical studies is determining an appropriate sample size. When more complex models are used, estimates of power can only be obtained using Monte Carlo simulations. In this tutorial, we introduce the R package `mlpwr` to perform simulation-based power analysis based on surrogate modeling. Surrogate modeling is a powerful tool to guide the search for study design parameters that imply a desired power or meet a cost threshold (e.g., in terms of monetary cost). `mlpwr` can be used to search for the optimal allocation when there are multiple design parameters, e.g., when balancing the number of participants and the number of groups in multilevel modeling. At the same time, the approach can take into account the cost of each design parameter, and aims to find a cost-efficient design. We introduce the basic functionality of the package, which can be applied to a wide range of statistical models and study designs. Additionally, we provide two examples based on empirical studies for illustration: one for sample size planning when using an item response theory model, and one for assigning the number of participants and the number of countries for a study using multilevel modeling.

Keywords: simulation, sample size, power analysis, machine learning

Introduction

Reliable testing of scientific hypotheses requires a sufficiently large sample size. A ubiquitous challenge in empirical research is that recruiting large samples is difficult due to resource constraints (e.g., time, money, labor) or ethical constraints (e.g., inconvenience or participation risks). However, if the sample sizes are small, random noise can mask the true effects, e.g. with regard to observed behaviour or cognitive processes. In a formal hypothesis testing framework, this trade-off between resource constraints and statistical significance is best described by the measure of statistical power. Statistical power describes the probability of finding an effect that is actually present in the population of interest. In general, we want our sample size to be large enough to achieve high statistical power while using as few resources as necessary.

Justifying Sample Sizes

The recent replication crisis has put low statistical power and replicability of scientific research into focus (Button et al., 2013; Open Science Collaboration, 2015). Starting from the observation that most published research results might be wrong (Ioannidis, 2005; Simmons et al., 2011), there have been several developments to improve the replicability of scientific studies (Shrout & Rodgers, 2018). One of these are registered reports, in which research projects are reviewed and conditionally accepted based on sound methodology rather than on the statistical significance of the result. In registered reports, justification of sam-

Felix Zimmer Mirka Henninger Rudolf Debelak

This material is based upon work supported by the Swiss National Science Foundation under Grant No. 188929 awarded to Rudolf Debelak.

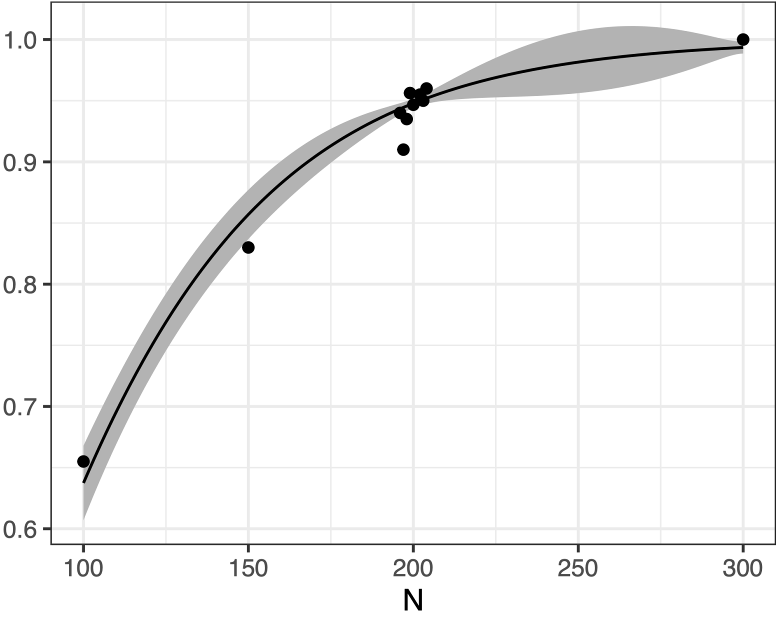
The R syntax for this study is available at the Open Science Framework at https://osf.io/xebsj/. All R packages used in this study are available on CRAN.

Correspondence concerning this article should be addressed to Felix Zimmer, Psychological Methods, Evaluation and Statistics, Department of Psychology, University of Zurich, Binzmuehlestrasse 14, Box 27, 8050 Zurich, Switzerland. E-mail: felix.zimmer@uzh.ch

Surrogate modeling

- Surrogate modeling aims to approximate a relationship that is costly to investigate with a cheaper function (Bhosekar Ierapetritou, 2018; Forrester Keane, 2009).
- We can adopt the idea of surrogate modeling to the functional relationship between study design parameters and power.
- Using this functional relationship, we can predict the power for a sample size that we did not perform a simulation at beforehand.
- Surrogate modeling is more efficient than grid search: In a simple example, our approach required only 20% of the computational effort and performed 50% more simulation runs that used the optimal sample size (Zimmer & Debelak, 2022).

Example



Slide explaining how we calculate the final sample size

The minimum sample that is within 10% of expect large sample performance in 80% of replications.

Example 1: Binary outcome, logistic regression

Example 2: Linear outcome, XGBoost

What’s next?

Package in development, functionining but more testing needed. Follow fediscience.org/@ewan for updates or LINK TO SIGN-UP.

1. Machine learning

2. Longitudinal data

3. Performance

Questions?

References I

[1] Laure Wynants et al. “Prediction Models for Diagnosis and Prognosis of Covid-19 Infection: Systematic Review and Critical Appraisal”. In: *BMJ* 369 (Apr. 2020). ISSN: 1756-1833. DOI: 10.1136/bmj.m1328. (Visited on 04/21/2020).

[2] Constanza L. Andaur Navarro et al. “Risk of Bias in Studies on Prediction Models Developed Using Supervised Machine Learning Techniques: Systematic Review”. In: *BMJ* 375 (Oct. 2021), n2281. ISSN: 1756-1833. DOI: 10.1136/bmj.n2281. (Visited on 07/31/2023).

[3] Paula Dhiman et al. “Methodological Conduct of Prognostic Prediction Models Developed Using Machine Learning in Oncology: A Systematic Review”. In: *BMC Medical Research Methodology* 22.1 (Apr. 2022), p. 101. ISSN: 1471-2288. DOI: 10.1186/s12874-022-01577-x. (Visited on 07/31/2023).

References II

[4] Alan J. Meehan et al. "Clinical Prediction Models in Psychiatry: A Systematic Review of Two Decades of Progress and Challenges". In: *Molecular Psychiatry* 27.6 (June 2022), pp. 2700–2708. ISSN: 1476-5578. DOI: 10.1038/s41380-022-01528-4. (Visited on 07/31/2023).