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

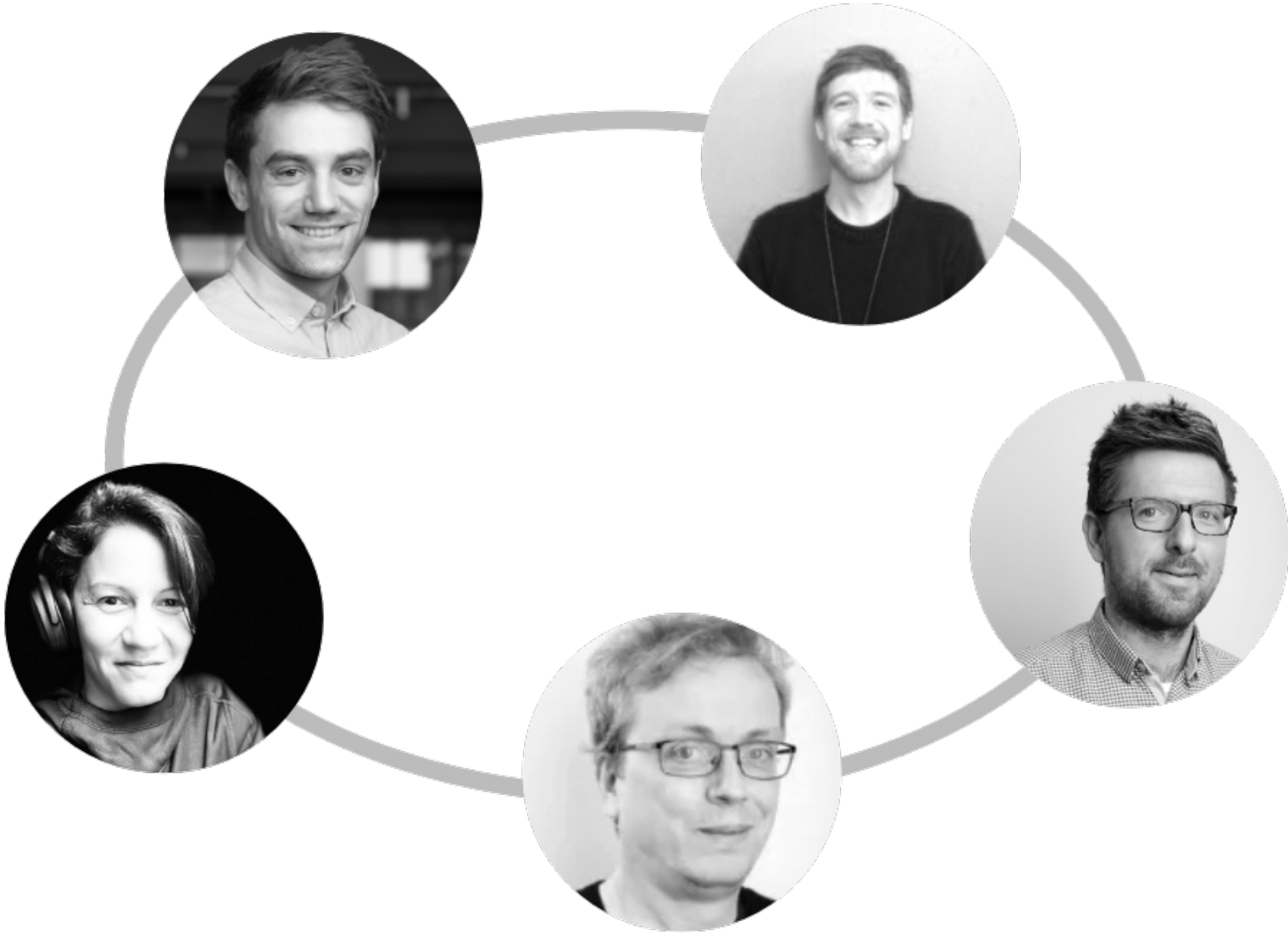
Department of Biostatistics & Health Informatics
King's College London

29th August 2023

30-second version

1. Most prediction models use small samples.
2. Small samples cause overfitting and imprecise estimates.
3. Existing tools can estimate minimum samples for continuous, binary, and survival outcomes.
4. Nothing exists for other models or data types.

We're developing a simulation-based approach that works with any outcome or method.



This talk

1. Background
   - What's the problem we're trying to solve?
   - What solutions currently exist?
2. Our simulation-based approach
   - Workflow and user interface
   - How it compares to other packages
3. Demonstration
4. Development status and next steps

We’re still developing the package.
Your feedback is welcome.
Please get in touch.

What's the problem?

Hundreds of prediction models are developed each year. Most have inadequate samples.

- Insufficient sample sizes was the most common cause of bias in 731 models for COVID-19.[2]
- Inadequate samples were found in:

67% models for COVID-19[2]

56% models using supervised machine learning[3]

73% models in psychiatry[5]

- Just 8% of machine learning models in oncology reported a sample size justification.[4]

Inadequate samples = research waste

- Inadequate samples lead to overfitting and inaccurate estimates of model parameters.
- This 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.

What tools currently exist?

Most studies ignore sample size.

Or use rules of thumb (e.g., 10 events per variable) that have no rationale in prediction modelling.[1]

In 2018, Riley et al released `pmsampsize` for R and Stata.

pmsampsize

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 has methods for simple continuous, binary, and survival outcome.

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, but. . .

We increasingly need to estimate minimum samples for:

Other models

- Regularised regression (e.g., LASSO, elastic net)
- Machine learning algorithms (e.g., random forests, gradient boosting)

Other types of data

- Longitudinal and repeated measures
- Clustered data

We're creating a simulation-based framework to estimate sample sizes for prediction.

The pmsims package for R

Key features:

- Able to estimate minimum sample sizes for any model or data type;
- Provides defaults for common model and data types;
- Efficient estimation.

This last point is key: most machine learning approaches are too computationally demanding for conventional simulation approaches.

Our approach

Setting

1. A study population represented by outcome-related individual characteristics (i.e., candidate predictors).
2. A chosen statistical or machine learning model.
3. Expected achievable performance (e.g., R^2, AUC) without sample size constraints, P^*.
4. Minimum acceptable performance of the model, P^{OK}.

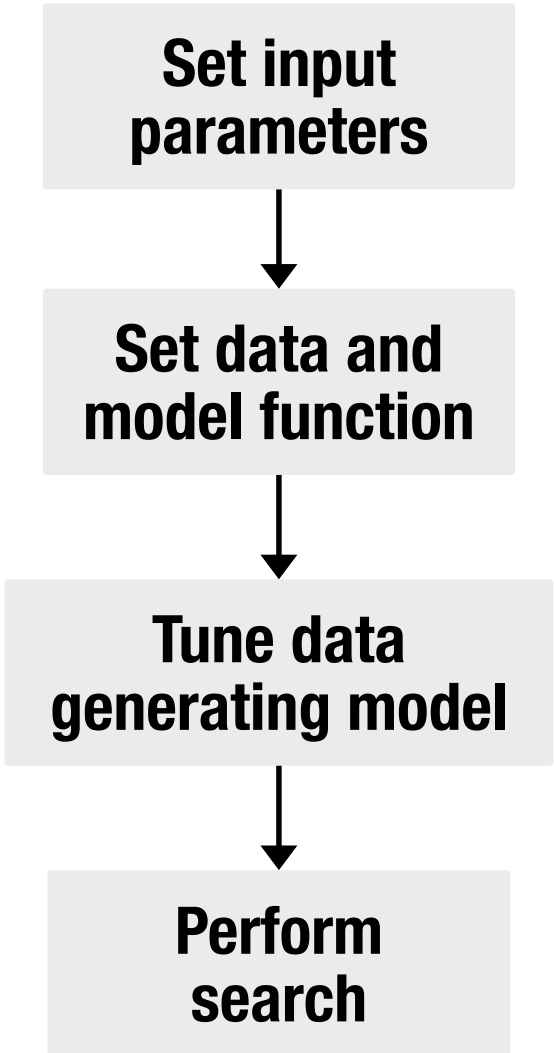
Our approach

Setting

1. A study population represented by outcome-related individual characteristics (i.e., candidate predictors).
2. A chosen statistical or machine learning model.
3. Expected achievable performance (e.g., R^2, AUC) without sample size constraints, P^*.
4. Minimum acceptable performance of the model, P^{OK}.

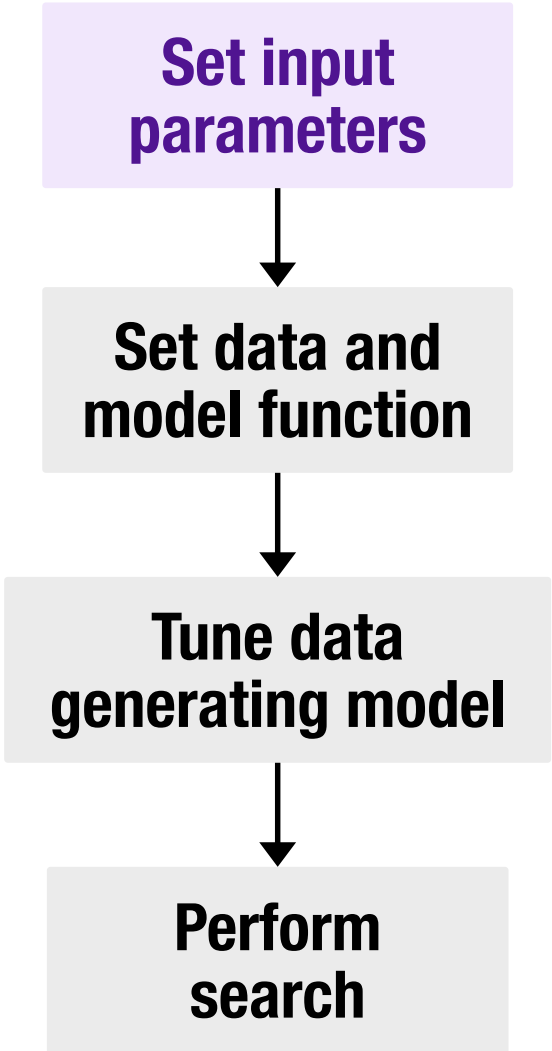
Find the minimum sample that ensures test performance of P^{OK} with probability of 80%, given the population, predictors, and P^*.

How does it work?



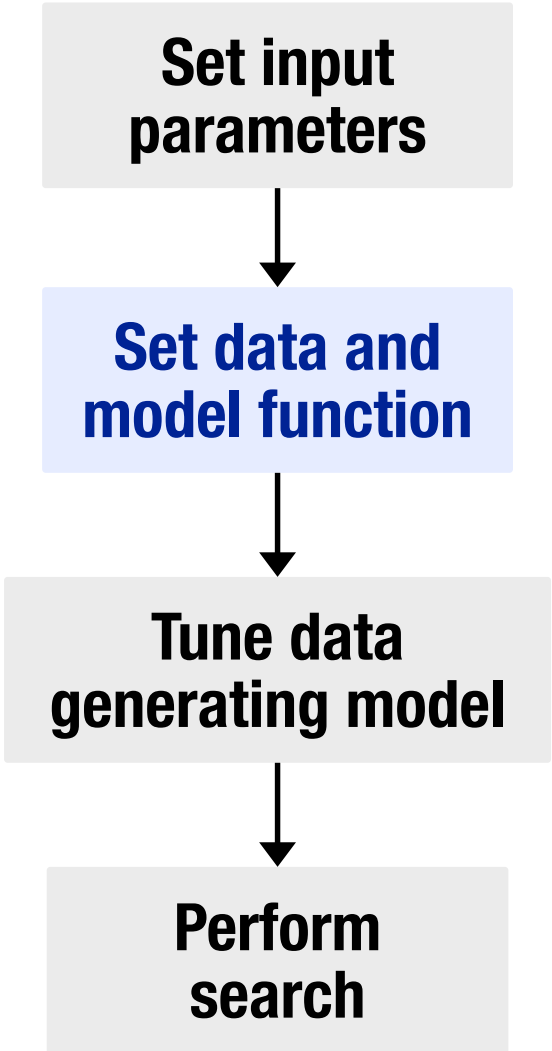
Some text here.

How does it work?



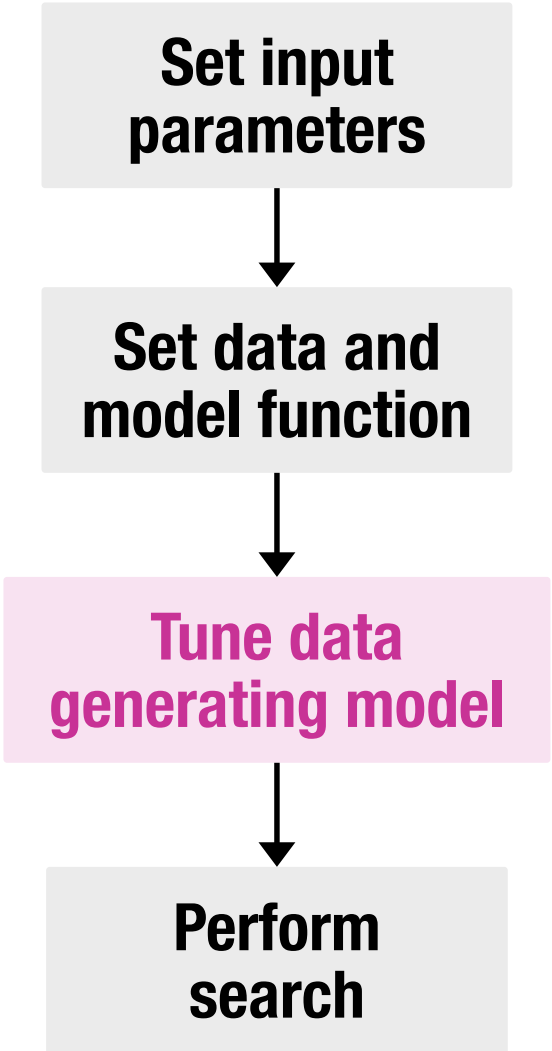
Some text here.

How does it work?



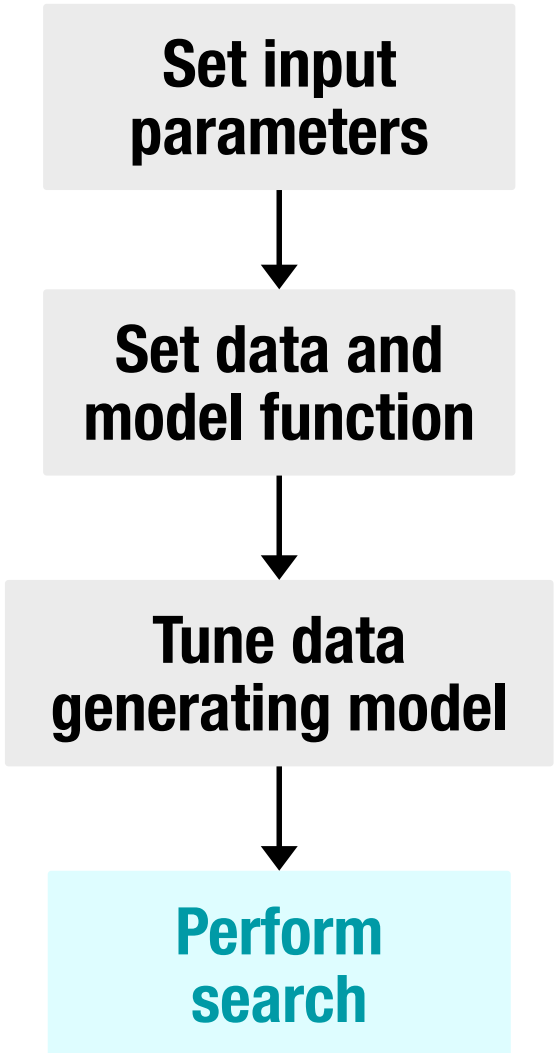
Some text here.

How does it work?



Some text here.

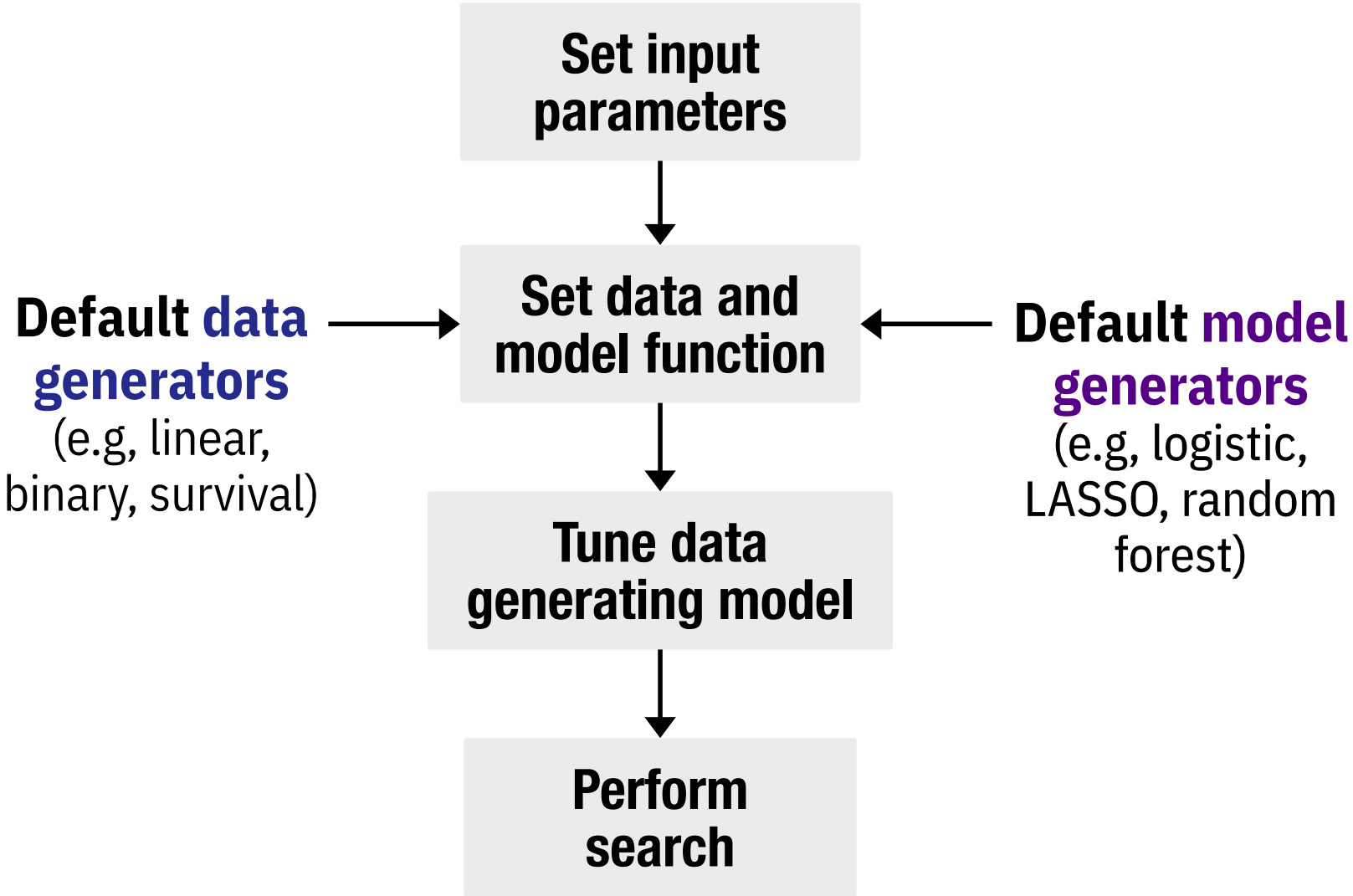
How does it work?



Some text here.

Maybe put slide here about conceptual differences vs. pmsampsize?

Our approach



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.

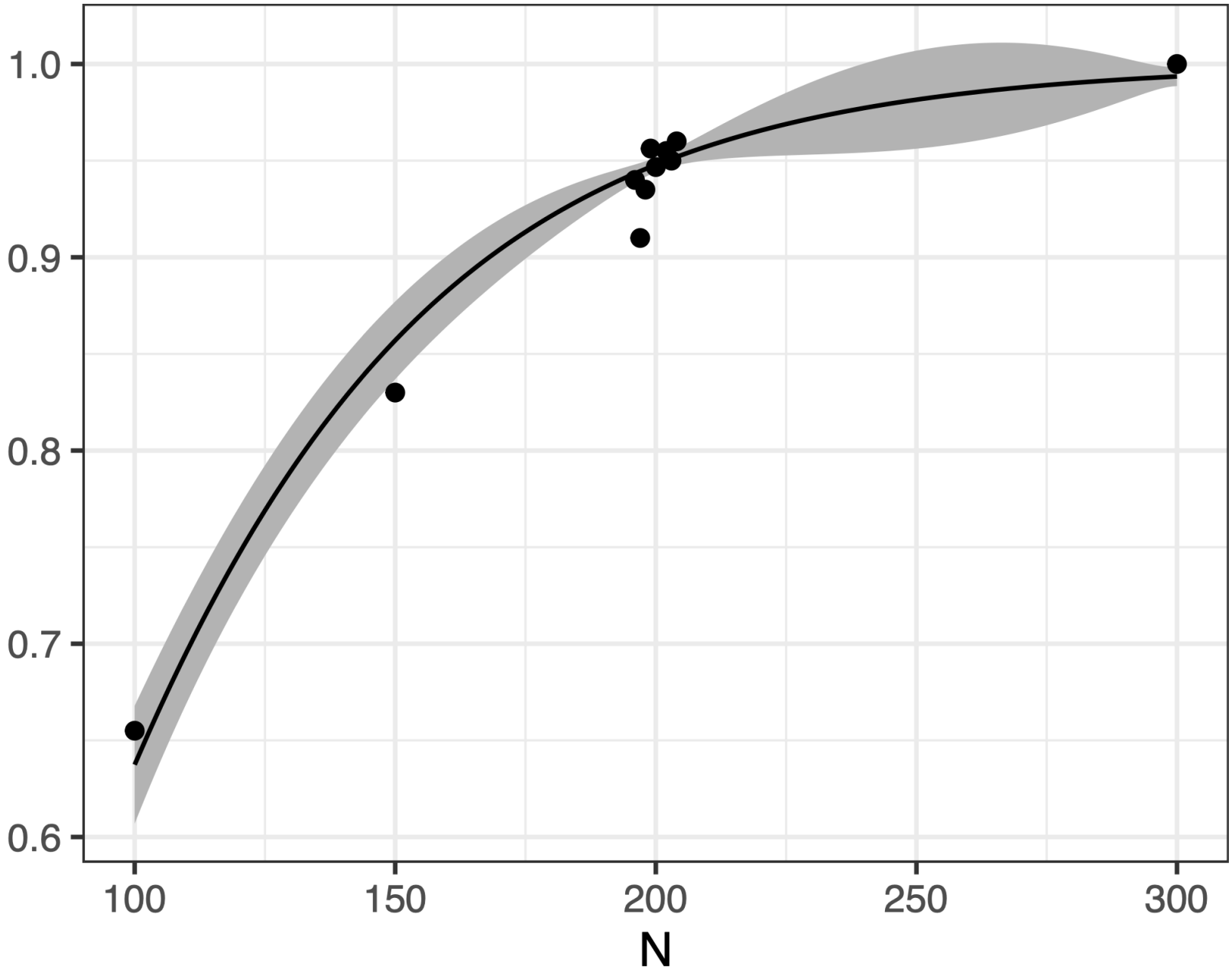
The R syntax for this study is available at the Open Science Framework at https://osf.io/xebsj/. All R packages used in this study are available on CRAN.

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

1. User specifies input parameters

   - The expected large sample performance of the model.
   - The range of sample sizes over which to search.
   - The number of signal and noise parameters.
   - The expected outcome prevalence.

2. Set data, model, and metric functions based on user input

   - Use defaults, but can be specified (e.g. `model = "lasso"`).

3. Tune the data generating model
4. Perform search; return minimum sample meeting criteria.

CRITERIA

We then return the minimum sample that is within 10% of expect large sample performance in 80% of replications.
If we had unlimited data, what's possible? Best case. What is the sample size that is sufficient to be within 10% of this maximum achieveable.
How many replications?

Example 1: Binary outcome, logistic regression

including comparison with `pmsampsize`

Example 2: Binary outcome, LASSO regression

Example 3: Custom model function

`simulate_custom`

XGBoost

Maybe put slide with simulations here?

With our package we can replicate other criteria.
How would it look if we included pmsampsize criteria within pmsims framework?
For example, shrinkage from `pmsampsize`:
DEMO
We can accomodate any.

Development status

Package in development; functioning, but more testing needed.

- Follow fediscience.org/@ewan for updates.
- Or enter an email address at tinyurl.com/is-it-ready-yet to get one email when a public release is available.

Please come and talk to us.
Criteria/models/etc. all subject to change.

What's next? (1/2)

1. Machine learning

2. Longitudinal data

What's next? (2/2)

3. Common data types

e.g., clinical, NLP, genetic.

4. Performance

Thank you for listening.

References I

[1] Maarten van Smeden et al. “No Rationale for 1 Variable per 10 Events Criterion for Binary Logistic Regression Analysis”. In: *BMC Medical Research Methodology* 16.1 (Nov. 2016), p. 163. ISSN: 1471-2288. DOI: 10.1186/s12874-016-0267-3. (Visited on 07/31/2023).

[2] Laure Wynants et al. “Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal”. In: *BMJ* 369 (Apr. 2020), p. m1328. ISSN: 1756-1833. DOI: 10.1136/bmj.m1328. (Visited on 08/16/2023).

[3] 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).

[4] 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).

[5] 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).