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

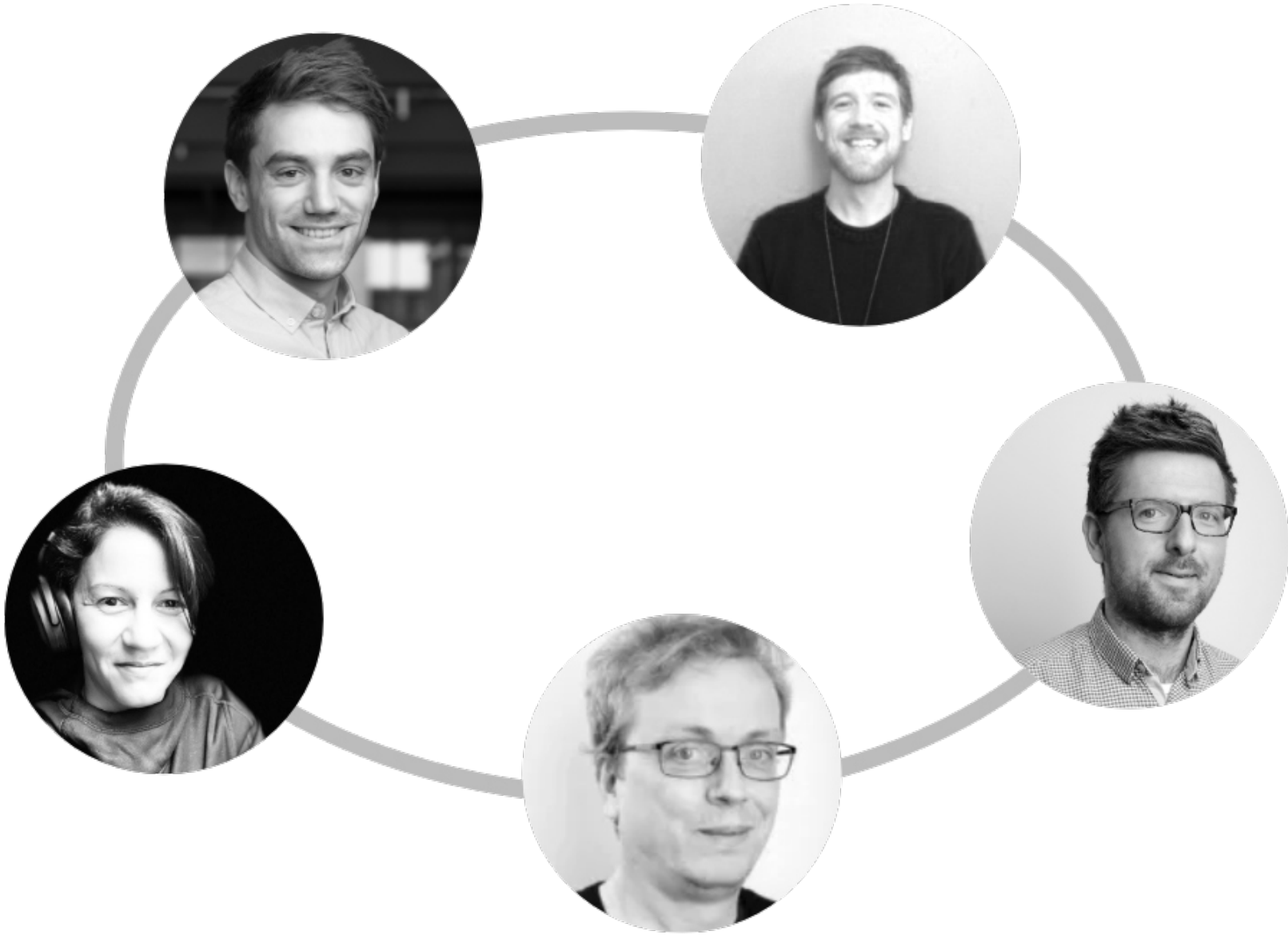
29th August 2023

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, with 67% (408/731) of models developed on too few patients (Wynants et al., 2020).
- 56% of models developed using supervised machine learning techniques are developed using inadequate sample sizes (Navarro et al., 2021).
- 73% of prediction models for psychiatry had inadequate sample size (Meehan et al., 2022).
- Only 8% of published machine learning models in oncology report a sample size justification.

Inadequate samples $\rightarrow$ 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 ≤ 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 model types, 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.

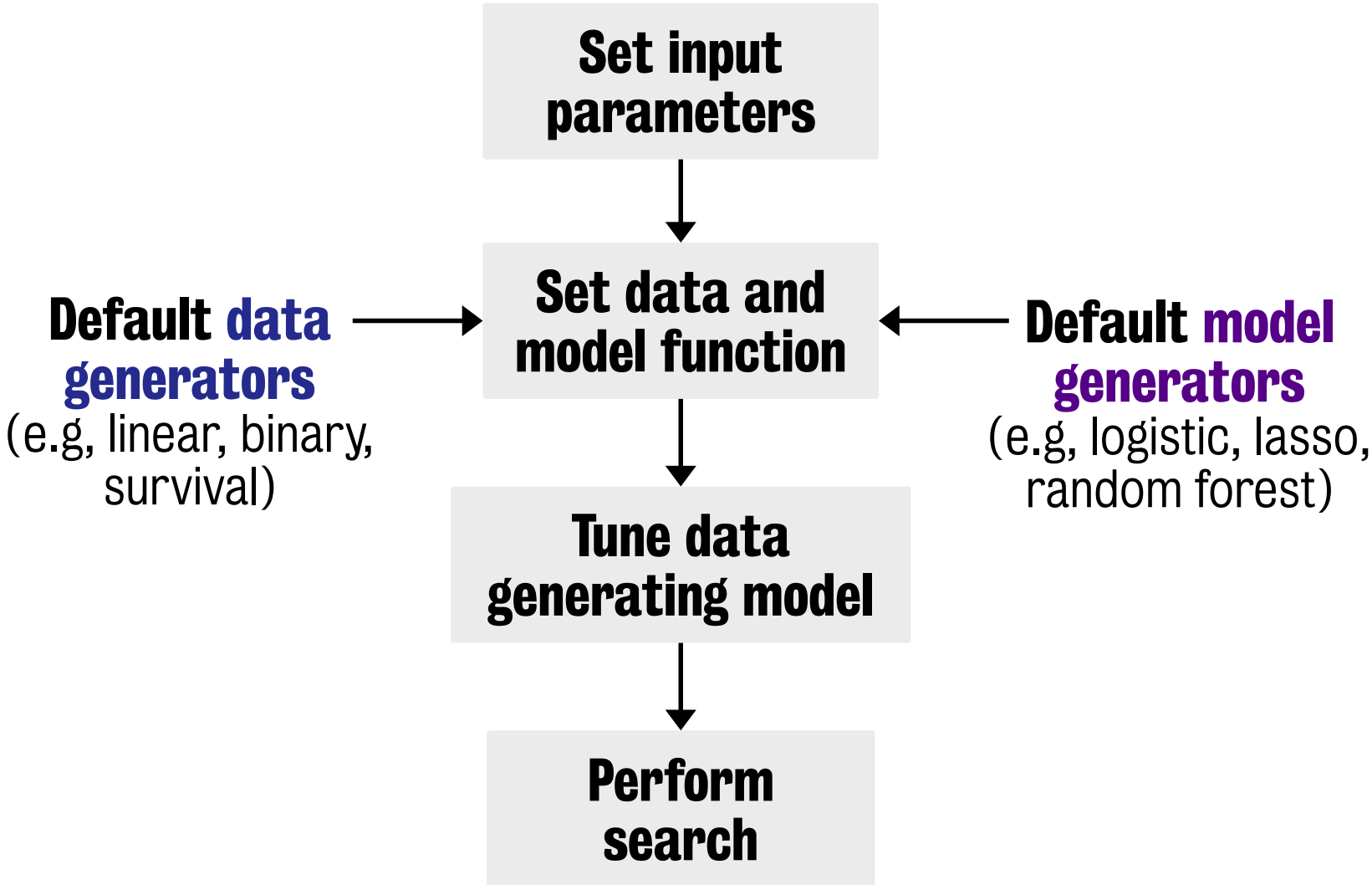
pmsims

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



Slide explaining input parameters and generators

Performing the search: `mlpwr`

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

mlpwr is a R package by

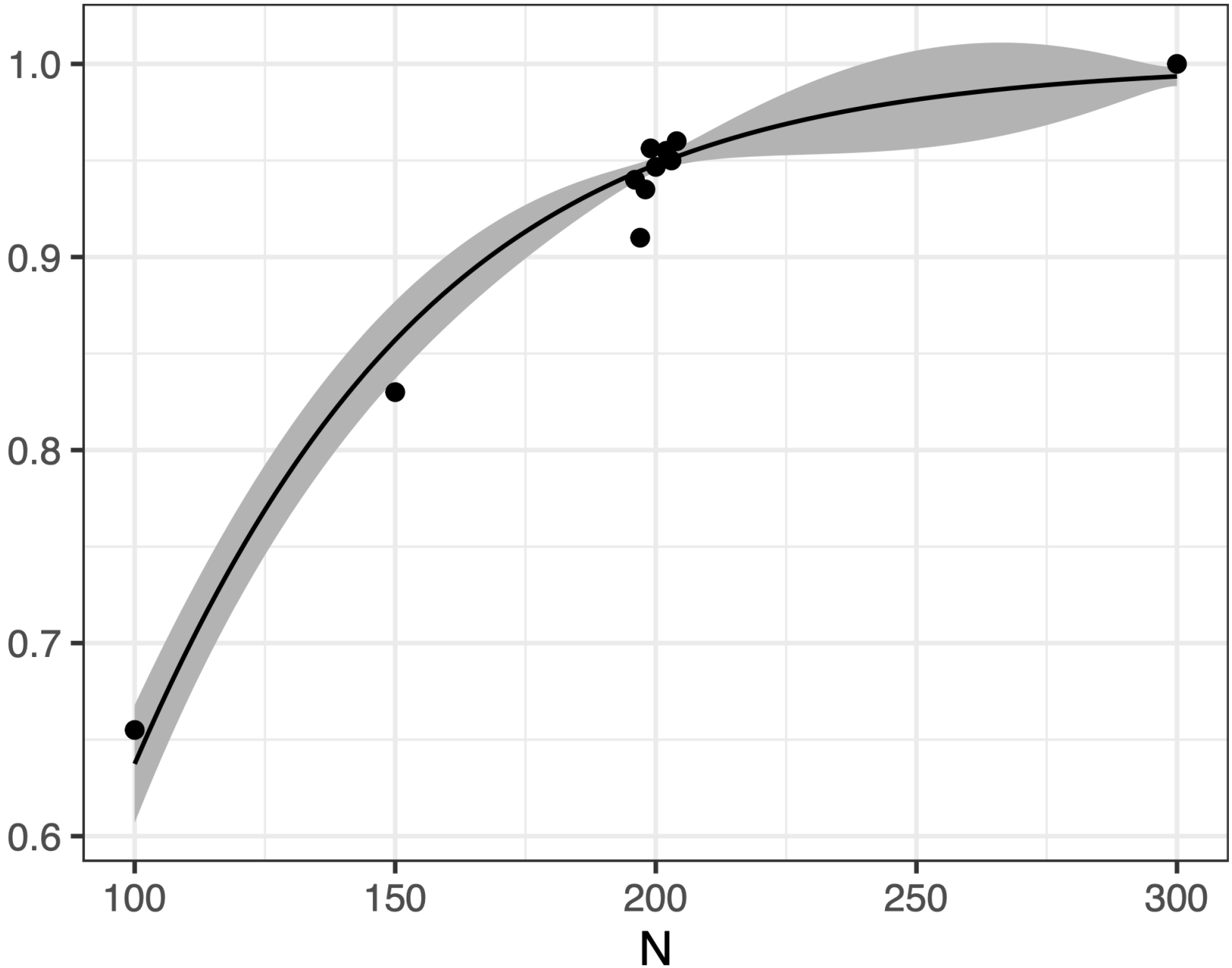
Felix Zimmer and Rudolf Debelak at the University of Zurich

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

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?