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

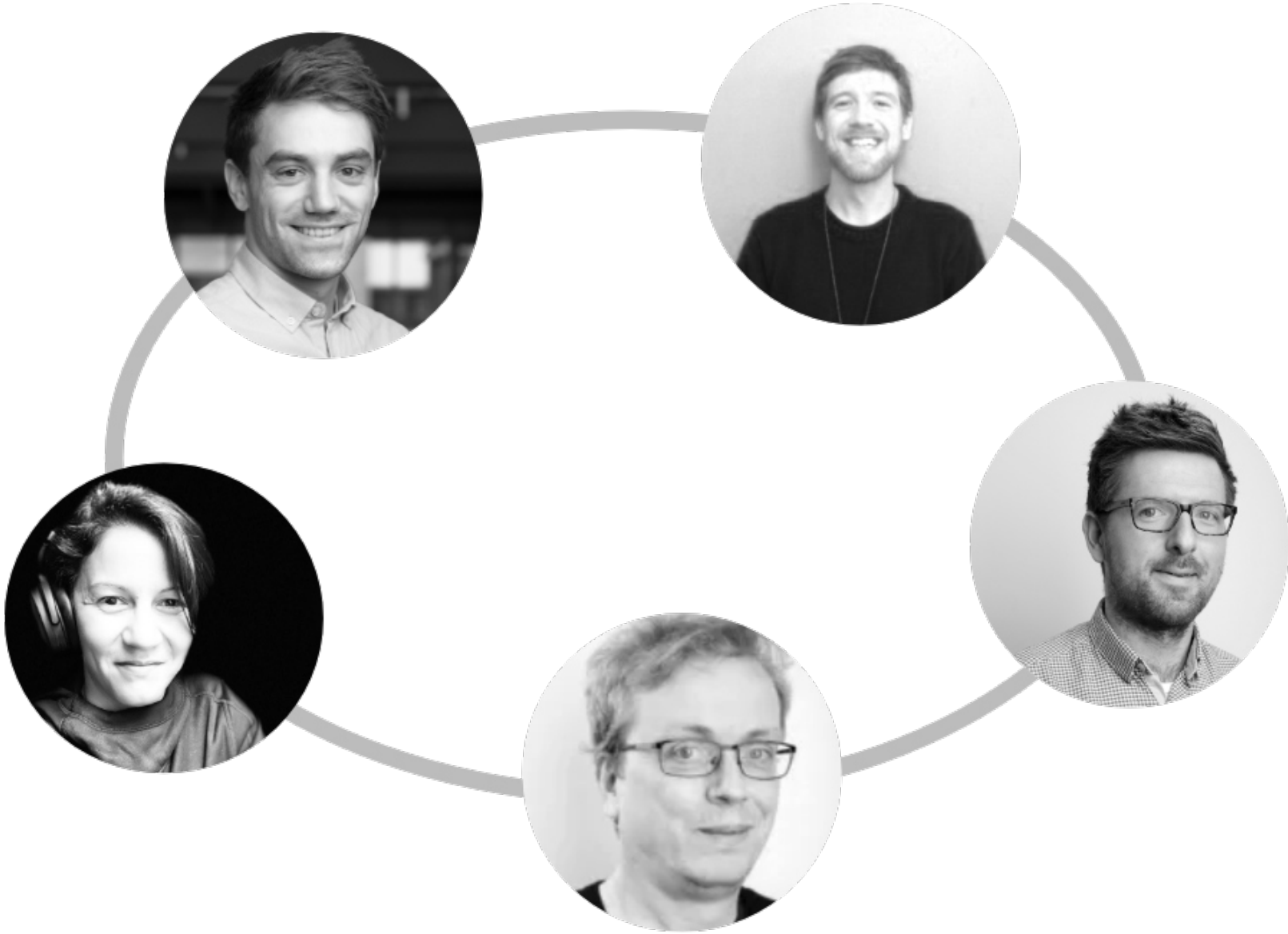
Department of Biostatistics & Health Informatics
King's College London

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.[3]
- Inadequate samples were found in:

67% models for COVID-19[3]

56% models using supervised machine learning[4]

73% models in psychiatry[6]

- Just 8% of machine learning models in oncology reported a sample size justification.[5]

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.[2]

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}.

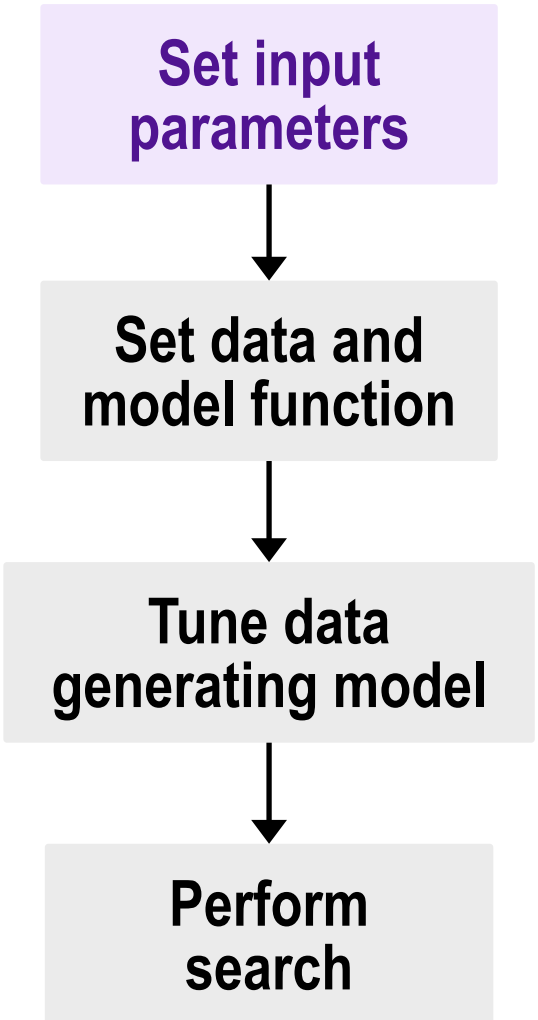
Our approach

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^*.

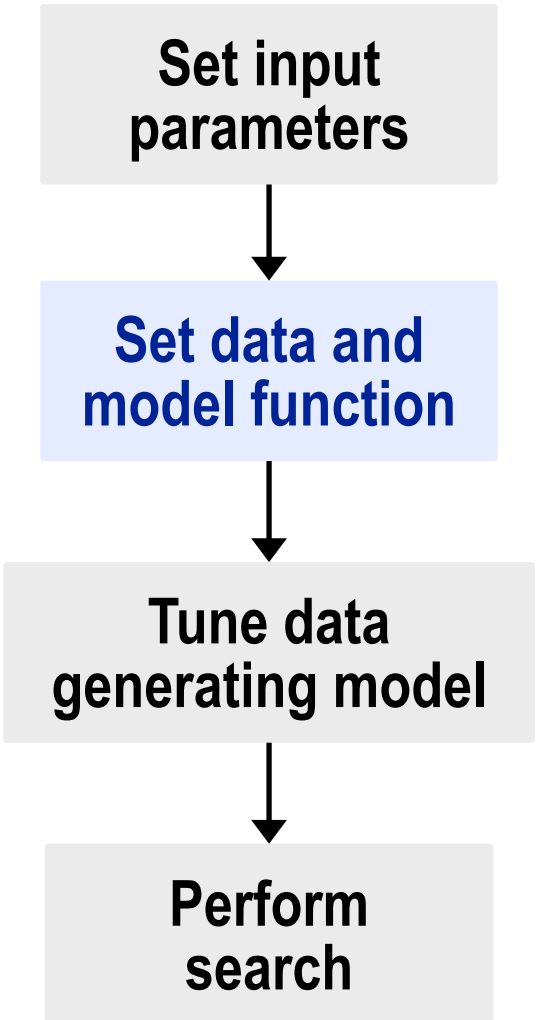
How does it work?



The user specifies:

1. The candidate predictors (number, type)
2. The chosen statistical model
3. The expected large sample performance (P^*)
4. The minimum acceptable performance (P^{OK})

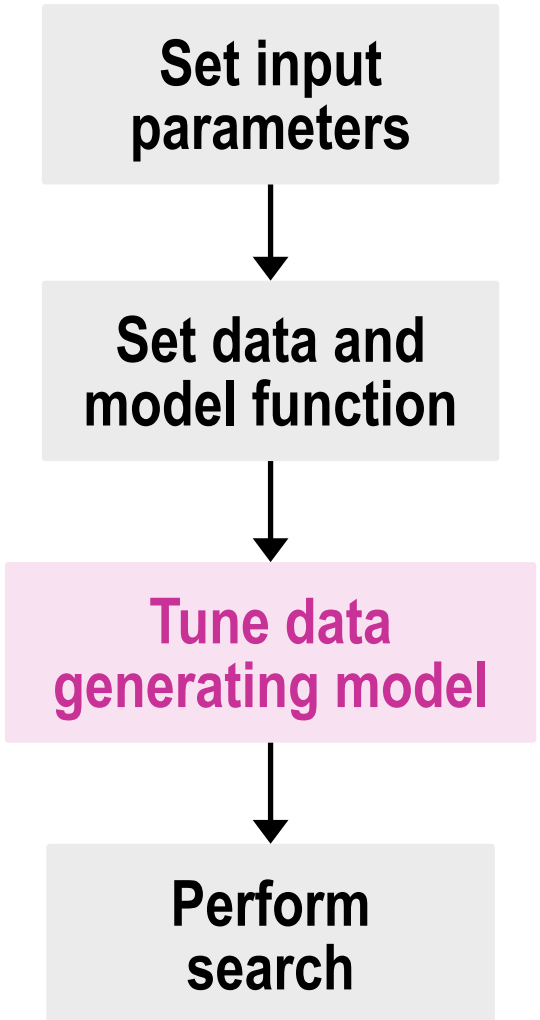
How does it work?



Based on their input, we set:

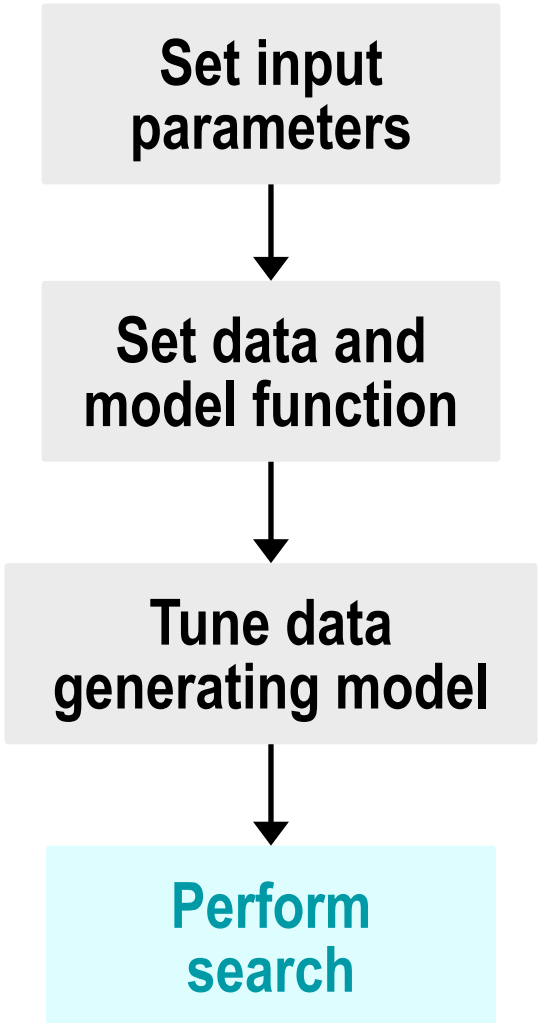
1. A data generating function
2. A model function
3. A metric function

How does it work?



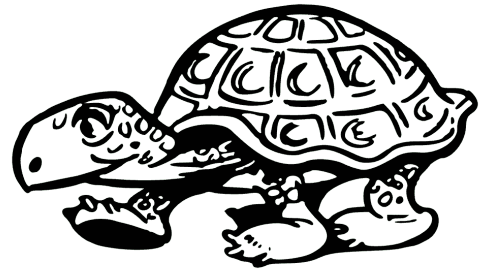
We then tune the data generating model, so the large sample performance is P^*.

How does it work?



Performing the search

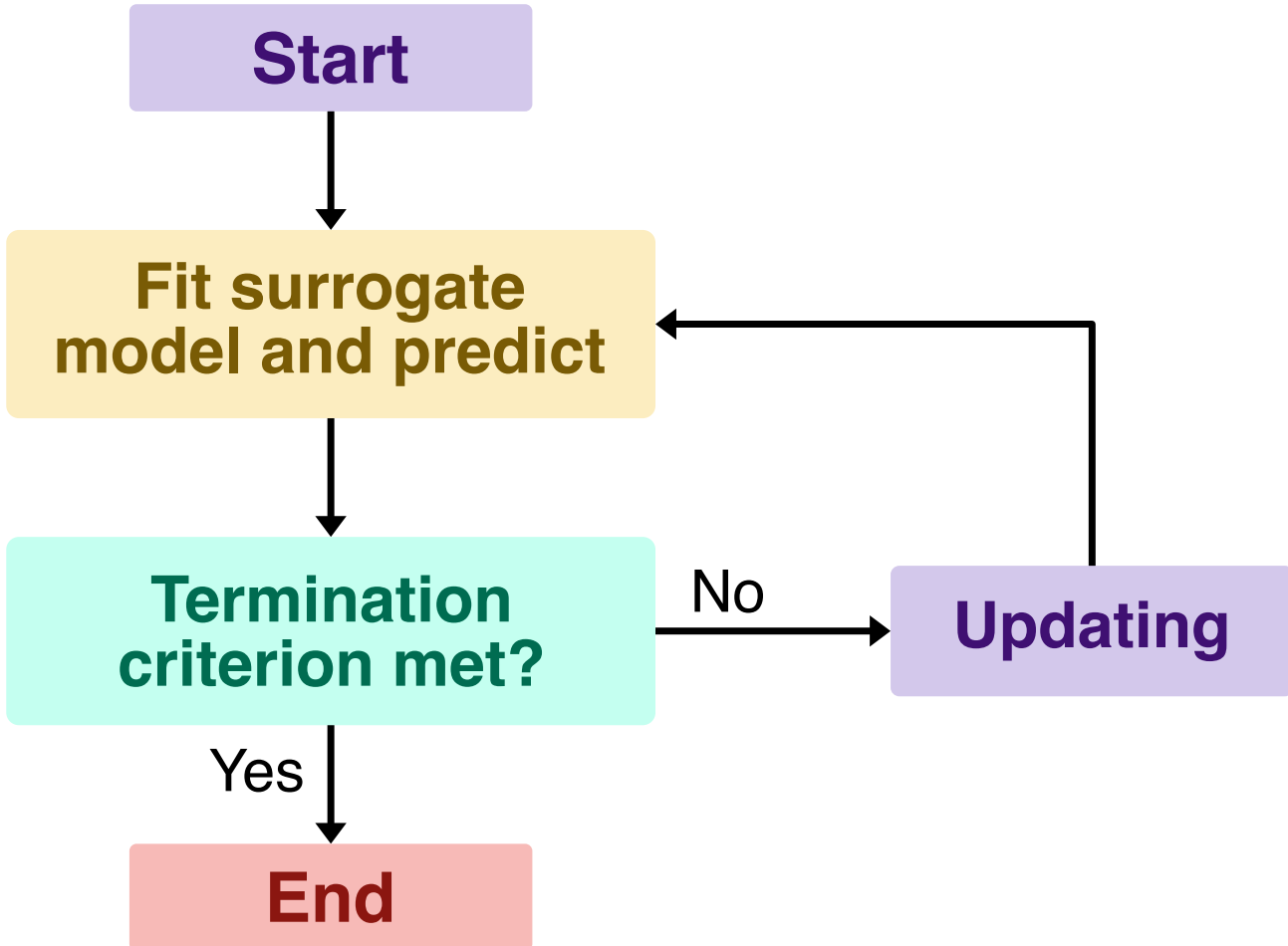
An exhaustive grid search would be too slow for many model types.



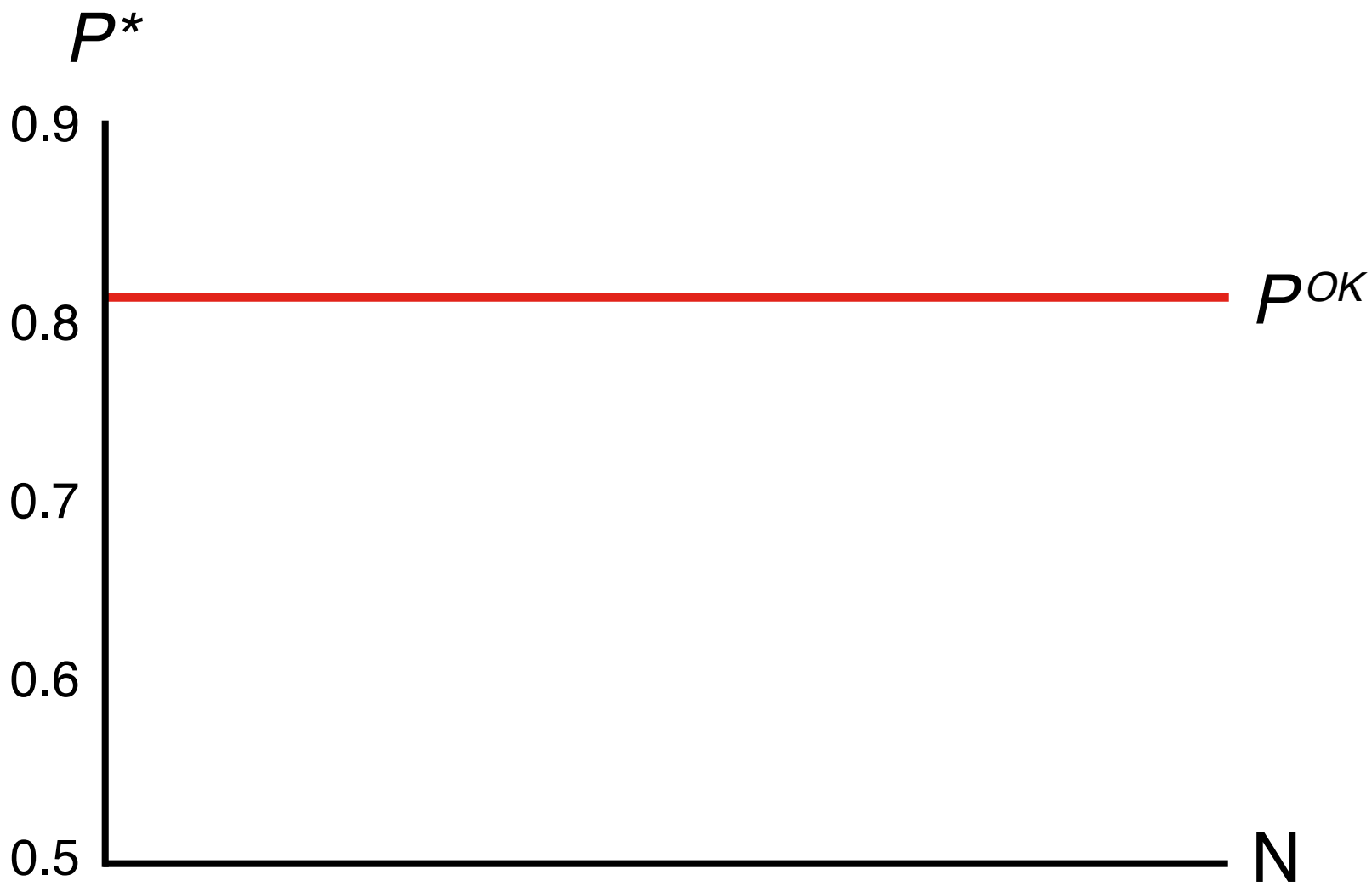
Surrogate modelling

- Surrogate modelling approximates the relationship between sample size and P^{OK}.
- Also referred to as 'learning curve fitting'.[1,7]
- We're building on the mlpwr R package by Felix Zimmer and Rudolf Debelak CITE. *Simulation-based Design Optimization for Statistical Power: Utilizing Machine Learning* (Forthcoming, JOURNAL)
- This uses Gaussian process regression.

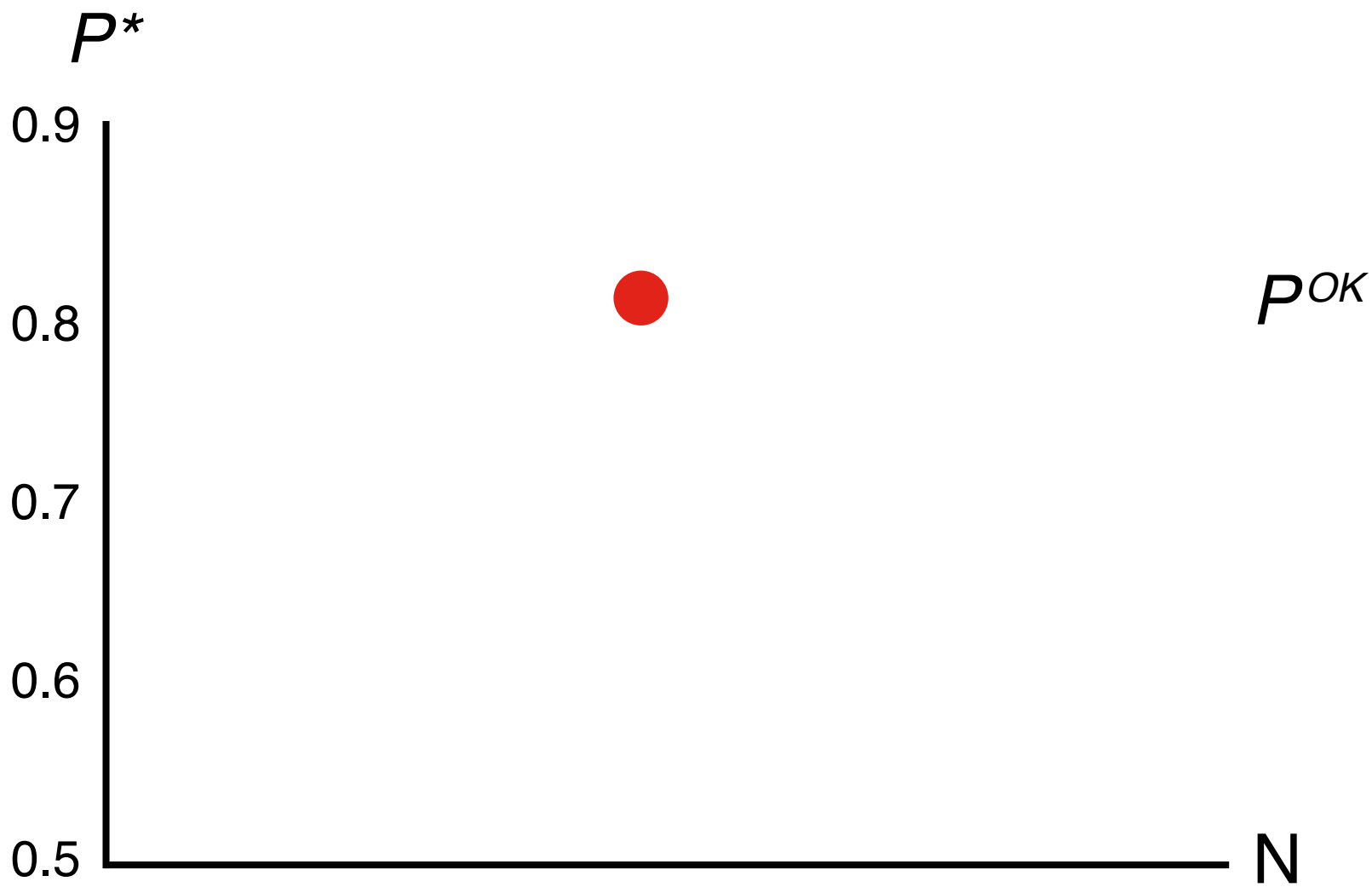
Surrogate modelling



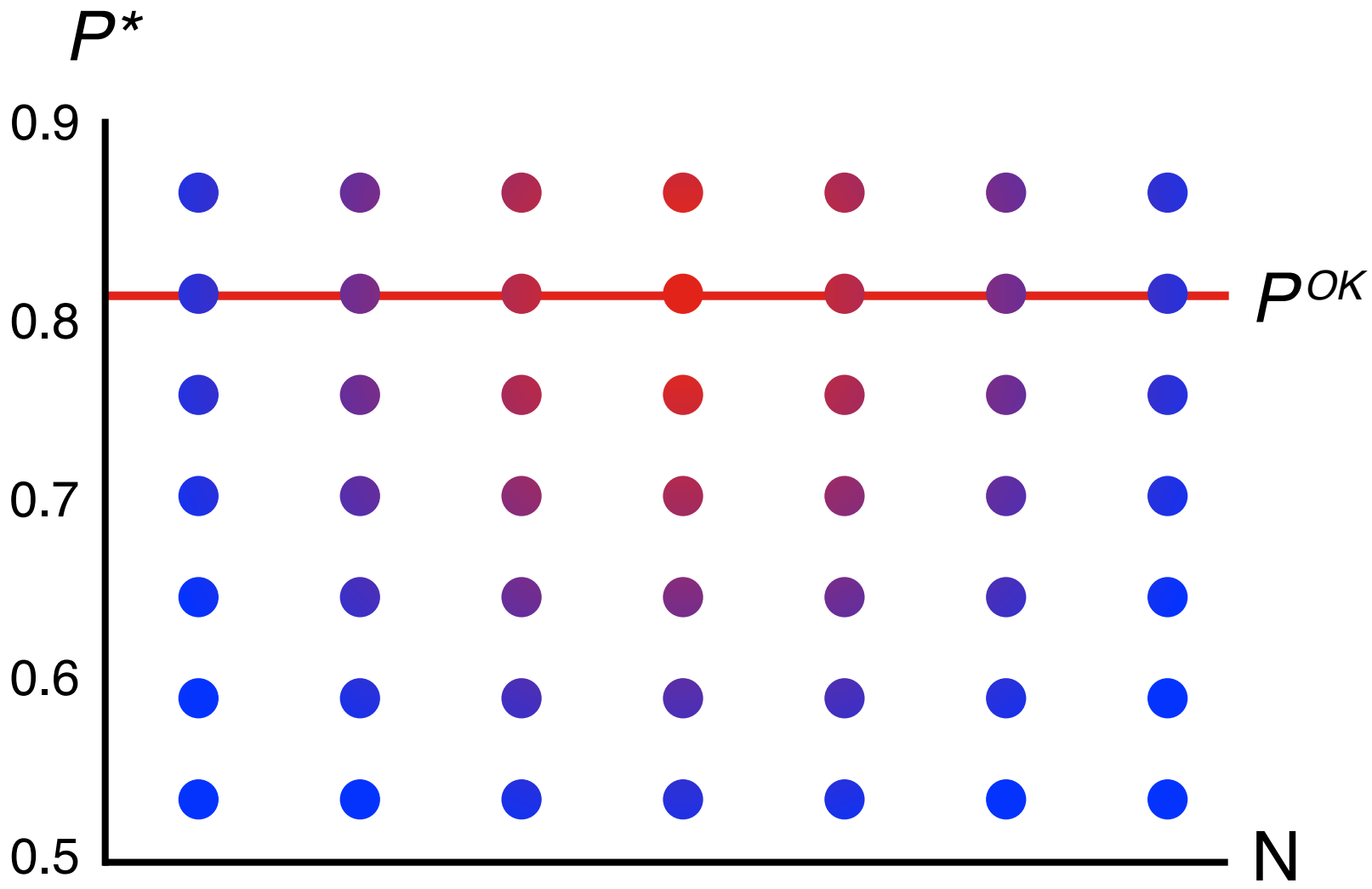
Surrogate modelling



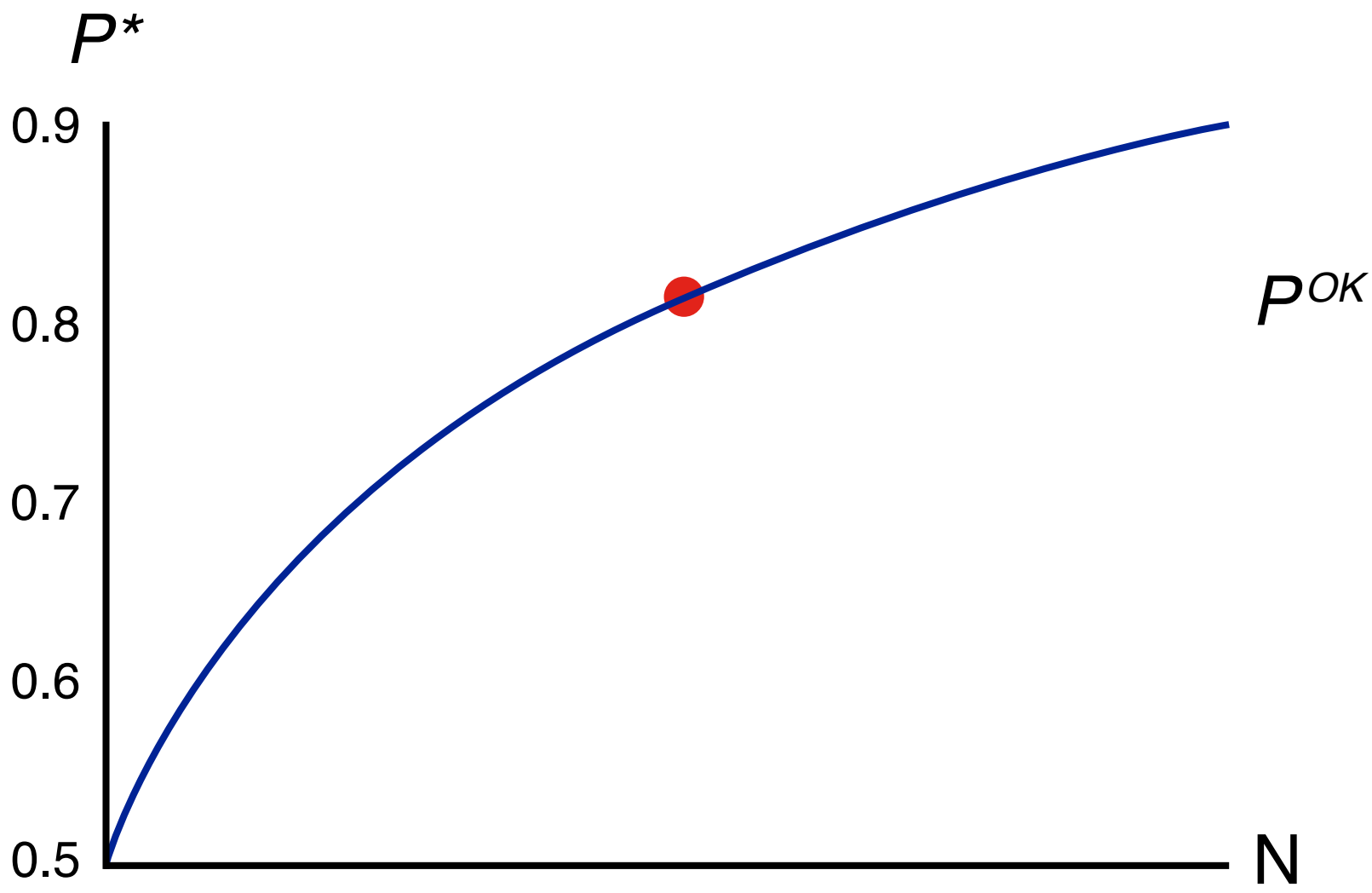
Surrogate modelling



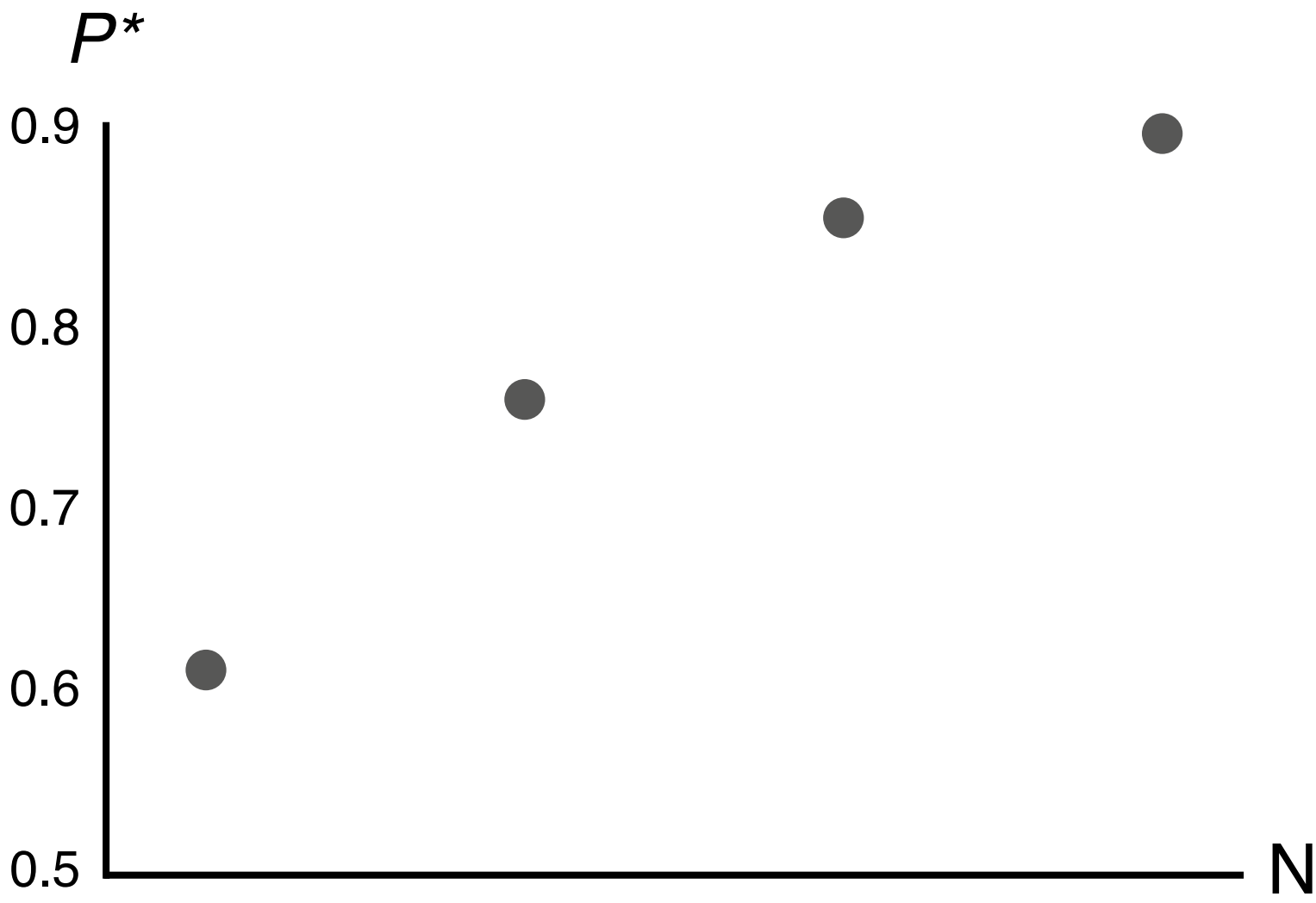
Surrogate modelling



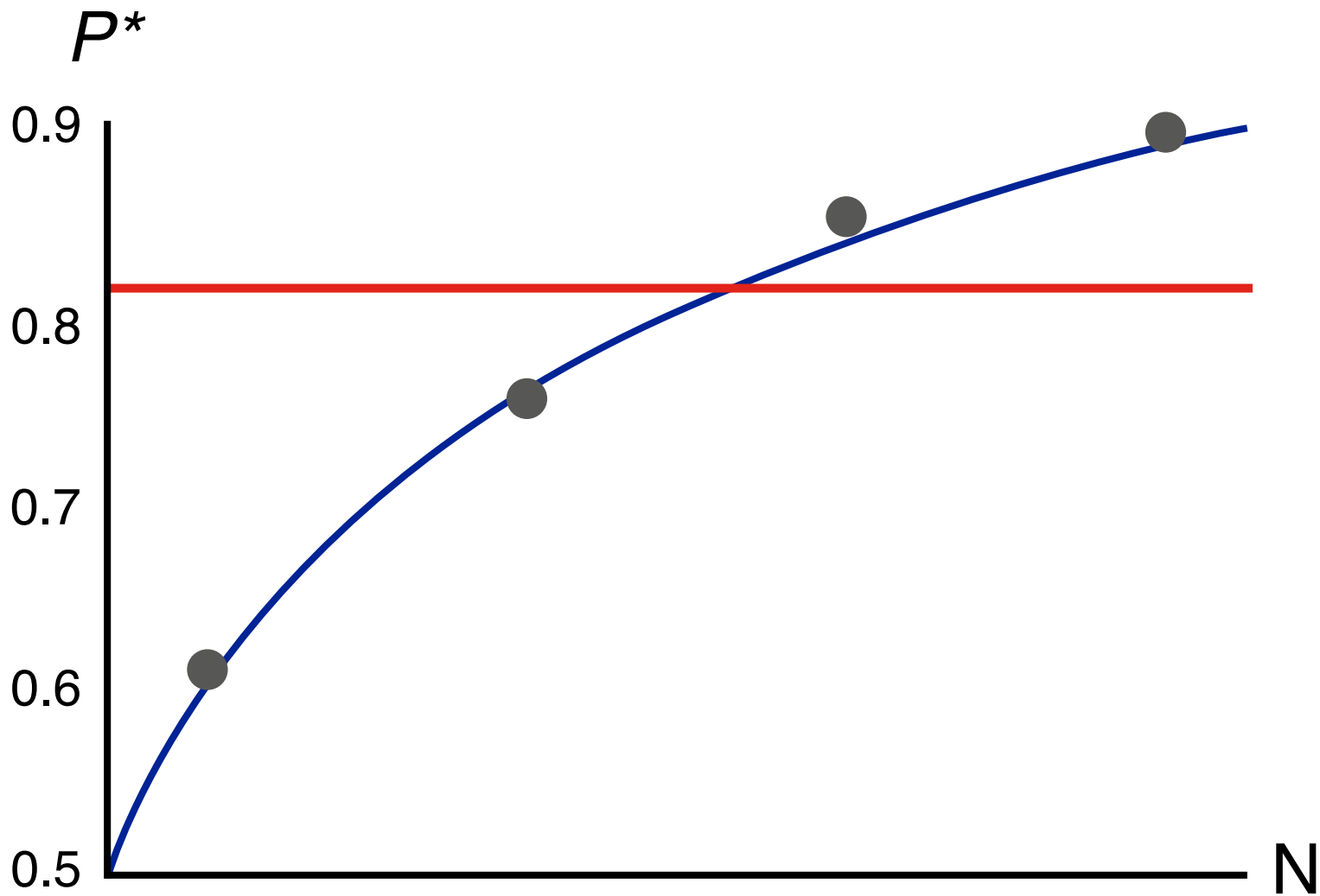
Surrogate modelling



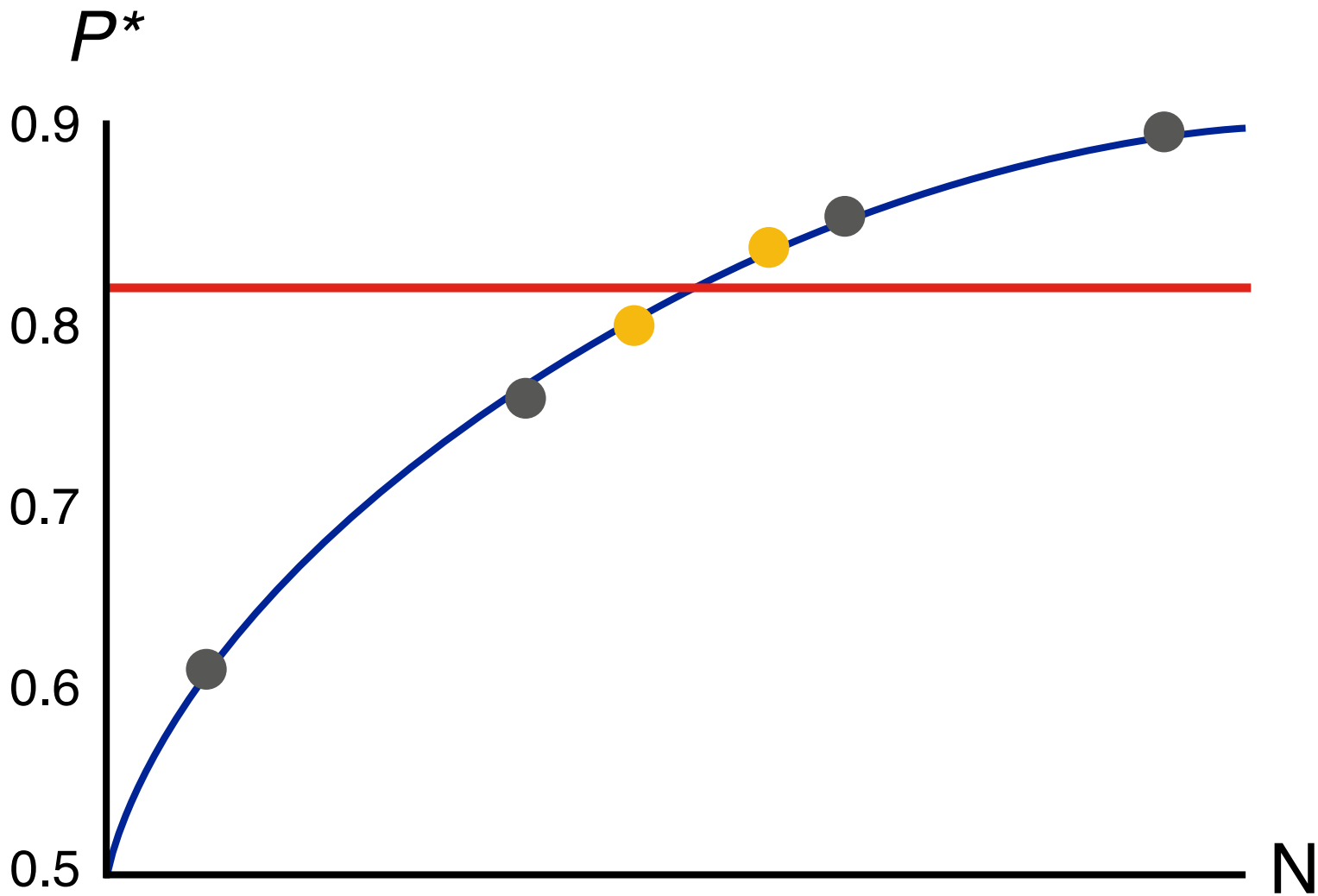
Surrogate modelling



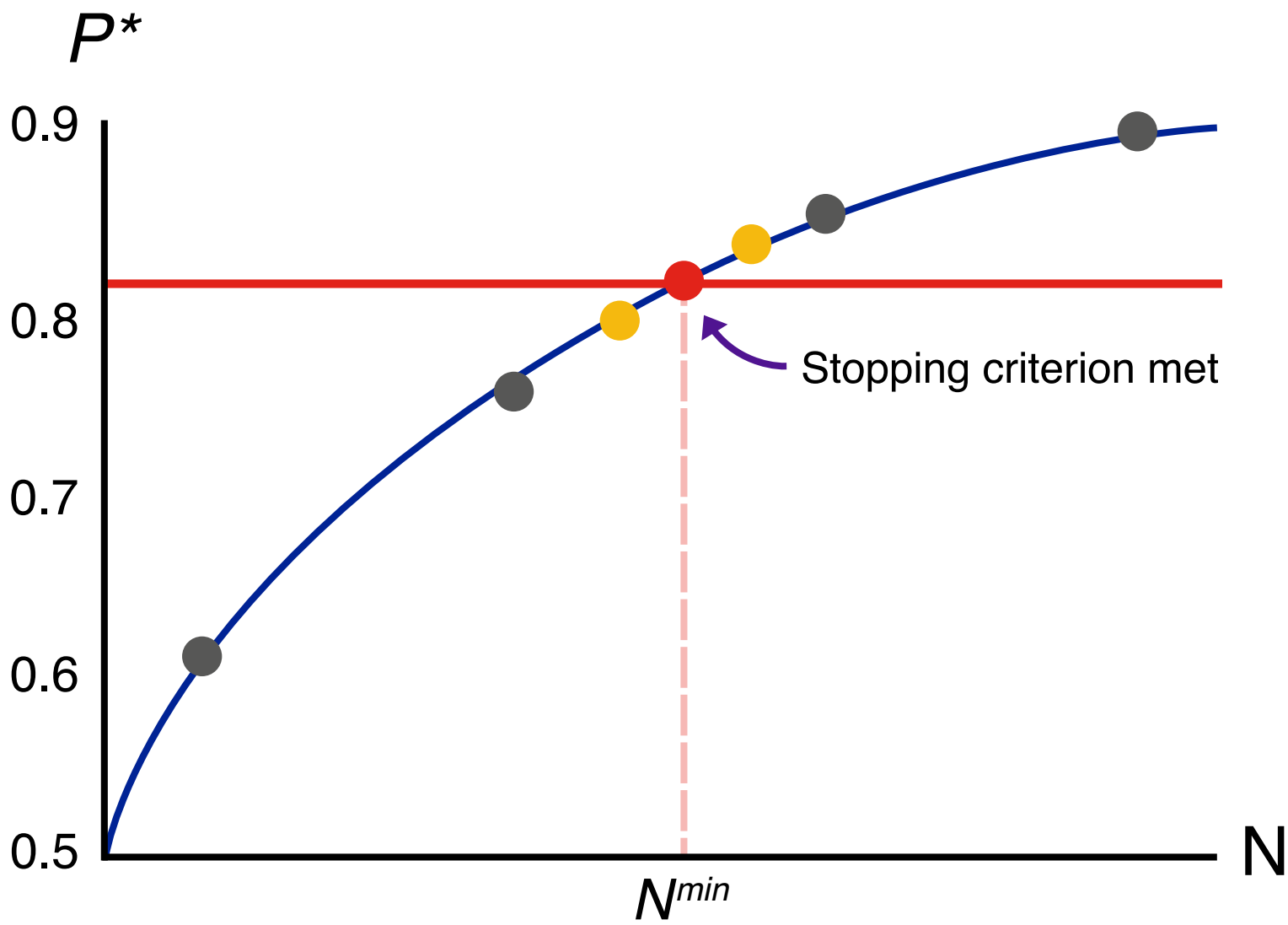
Surrogate modelling



Surrogate modelling



Surrogate modelling



How do we assess performance?

We identify the minimum sample that meets two criteria:

1. Discrimination

Within 0.1 of the achievable performance without sample size constraints, P^*

The choice of metric and metric can be set by users

```
minimum_threshold = 0.1 metric = "auc"
```

2. Calibration

Within 0.1 of a calibration slope of 1.0

Two approaches to estimating minimum samples

	pmsims	pmsampsize
Target?	Find training sample size such that an expected apparent and test performances are sufficiently close to each other	Find sample size such that with 80% probability test performance is above P^*
How?	Simulate	Uniform shrinkage

- Tune data generator to an expected achievable performance.
- Sample training data of different sizes, compute performance metrics. Using mlpwr find n at which 0.2 quantile of test performances achieves Pok.
- NB pmsims handles calibration slope as a performance measure, which is the same as uniform shrinkage, as slope is defined as minimizing the error

- Consider GLM models, where estimates depend on a linear predictor, $x^T\beta$, with $\beta-$ OLS/ML estimates from the training sample.
- Using $s \cdot x^T\beta$, instead of $x^T\beta$ may prevent overfitting and perform better on unseen cases.

pmsims

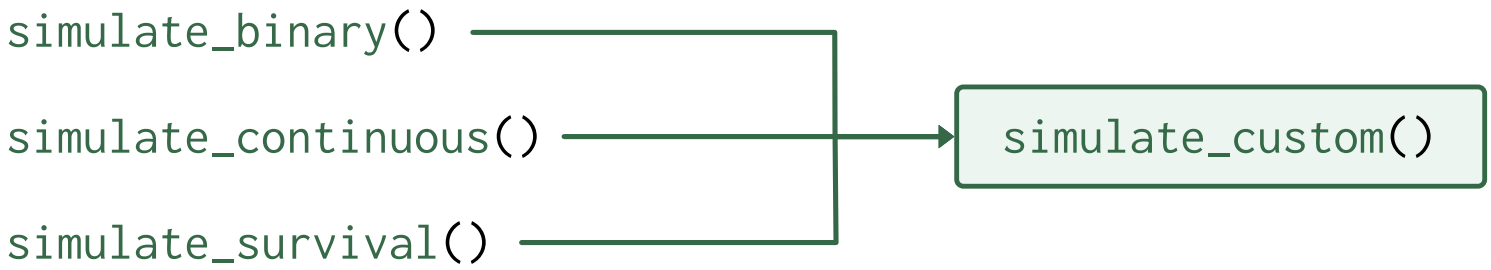
- Targets absolute performance
- Flexible, handles machine learning or multilevel models, etc.
- Does not aim to prevent overfitting per se
- Targets test performance itself *
- Adjusts recommendations to the test performance variance **

However:

- Takes time / computational resources
- Requires user input including the prediction model and/or data generator for complex designs
- Depends on simulations, so

pmsampsize

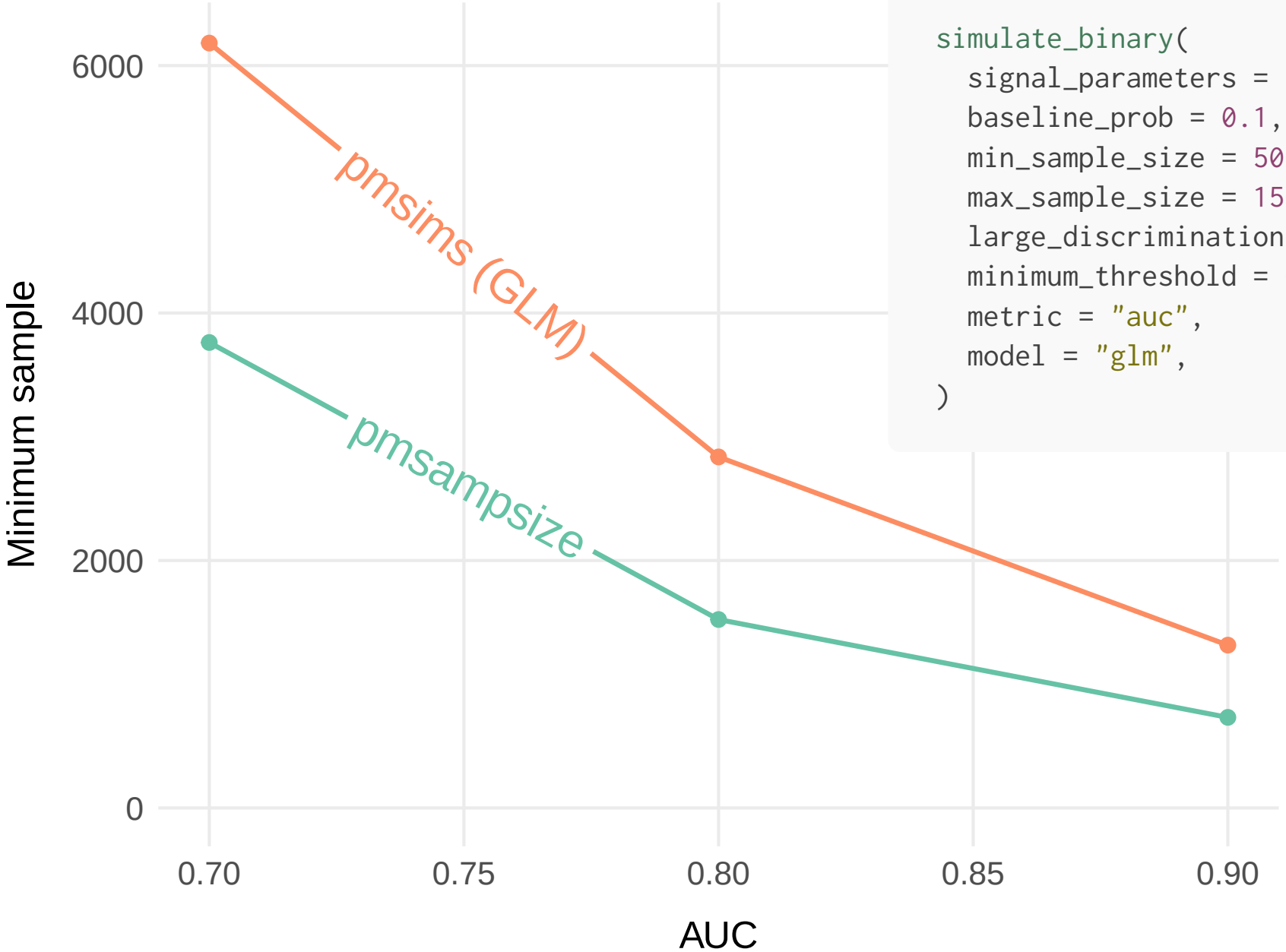
The user interface




```
simulate_continuous <-
  function(
    signal_parameters = 30,
    noise_parameters = 0,
    min_sample_size = 300,
    max_sample_size = 10000,
    large_discrimination = 0.7,
    minimum_threshold = 0.1,
    model = "lm",
    metric = "r2",
    ...
  )
```

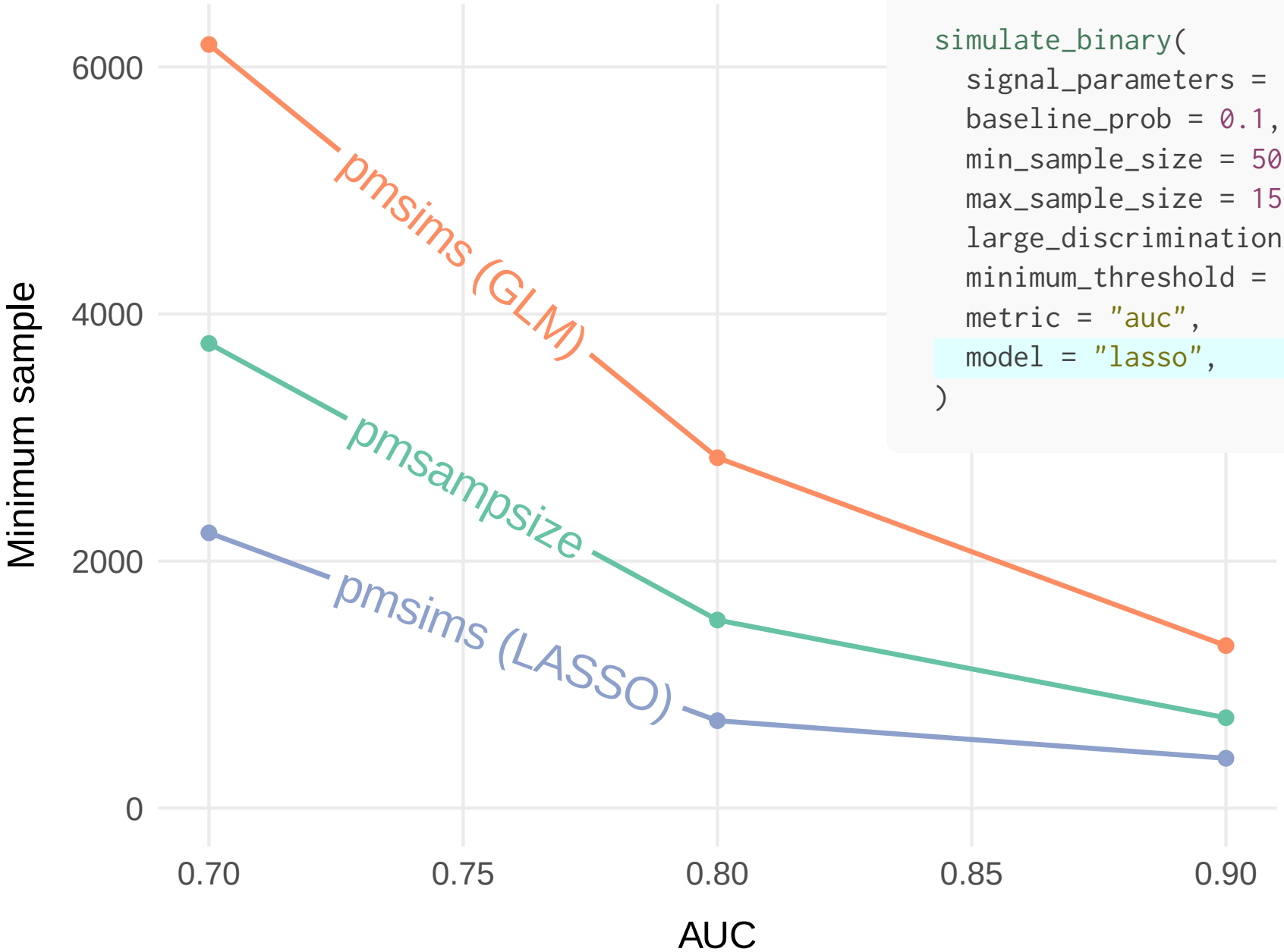
```
simulate_binary <-
  function(
    signal_parameters = 30,
    noise_parameters = 0,
    baseline_prob = 0.1,
    min_sample_size = 300,
    max_sample_size = 10000,
    large_discrimination = 0.8,
    minimum_threshold = 0.1,
    metric = "auc",
    model = "glm",
    ...
  )
```

Example 1: Binary outcome, logistic regression




```
simulate_binary(
  signal_parameters = 20,
  baseline_prob = 0.1,
  min_sample_size = 500,
  max_sample_size = 15000,
  large_discrimination = ...,
  minimum_threshold = 0.1,
  metric = "auc",
  model = "glm",
)
```

Example 2: Binary outcome, LASSO regression




```
simulate_binary(
  signal_parameters = 20,
  baseline_prob = 0.1,
  min_sample_size = 500,
  max_sample_size = 15000,
  large_discrimination = ...,
  minimum_threshold = 0.1,
  metric = "auc",
  model = "lasso",
)
```

Example 3: Custom model function

What if a model hasn't been implemented?

```
model_function <- function(d) {
  dmat <- xgboost::xgb.DMatrix(
    as.matrix(d[, -1]),
    label = d[, 1]
  )
  param <- list(
    objective = "binary:logistic",
    booster = "gblinear",
    alpha = 0.0001,
    lambda = 1
  )
  xgboost::xgb.train(
    param,
    dmat,
    nrounds = 2
  )
}
```

```
metric_function <- function(data,
                            fit,
                            model) {
  dmat <- xgboost::xgb.DMatrix(
    as.matrix(data[, -1]),
    label = data[, 1]
  )
  y_hat <- predict(fit, dmat)
  pROC::auc(data[, 1], y_hat)[1]
}
```

```
simulate_custom(
  data_function = data_function,
  model_function = model_function,
  metric_function = metric_function,
  ...
)
```

Development status

We're currently developing the package.

fediscience.org/@ewan for updates

ewan.carr@kcl.ac.uk

Enter email at tinyurl.com/is-pmsims-ready-yet to get one email when a public release is available.

Come and talk to us.

What's next?

1. Machine learning

2. Longitudinal data

3. Common data types

e.g., clinical, NLP, genetic.

4. Performance

Thank you for listening.

References I

[1] Rosa L. Figueroa et al. “Predicting Sample Size Required for Classification Performance”. In: *BMC Medical Informatics and Decision Making* 12.1 (Feb. 2012), p. 8. ISSN: 1472-6947. DOI: 10.1186/1472-6947-12-8. (Visited on 08/17/2023).

[2] 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).

[3] 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).

[4] 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).

[5] 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

[6] 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).

[7] Alimu Dayimu et al. *Sample Size Determination via Learning-Type Curves*. Mar. 2023. arXiv: 2303.09575 [stat]. (Visited on 08/16/2023).