4 kinds of stability:

Algorithmic Stability
Generalization
Hyperparameter Stability
Training stability

1. Algorithmic Stability

- This refers to how sensitive a machine learning algorithm is to changes in the training data.
- An algorithm is considered stable if small changes in the training data do not lead to significant changes in the model's predictions.
- Robust algorithms tend to be less sensitive to noise or outliers in the data.

2. Generalization

- A stable model should generalize well to unseen data.
- This means that the model should not overfit the training data but should instead capture the underlying patterns that are present in the data.
- A model that generalizes well is more likely to perform consistently on new, unseen data.

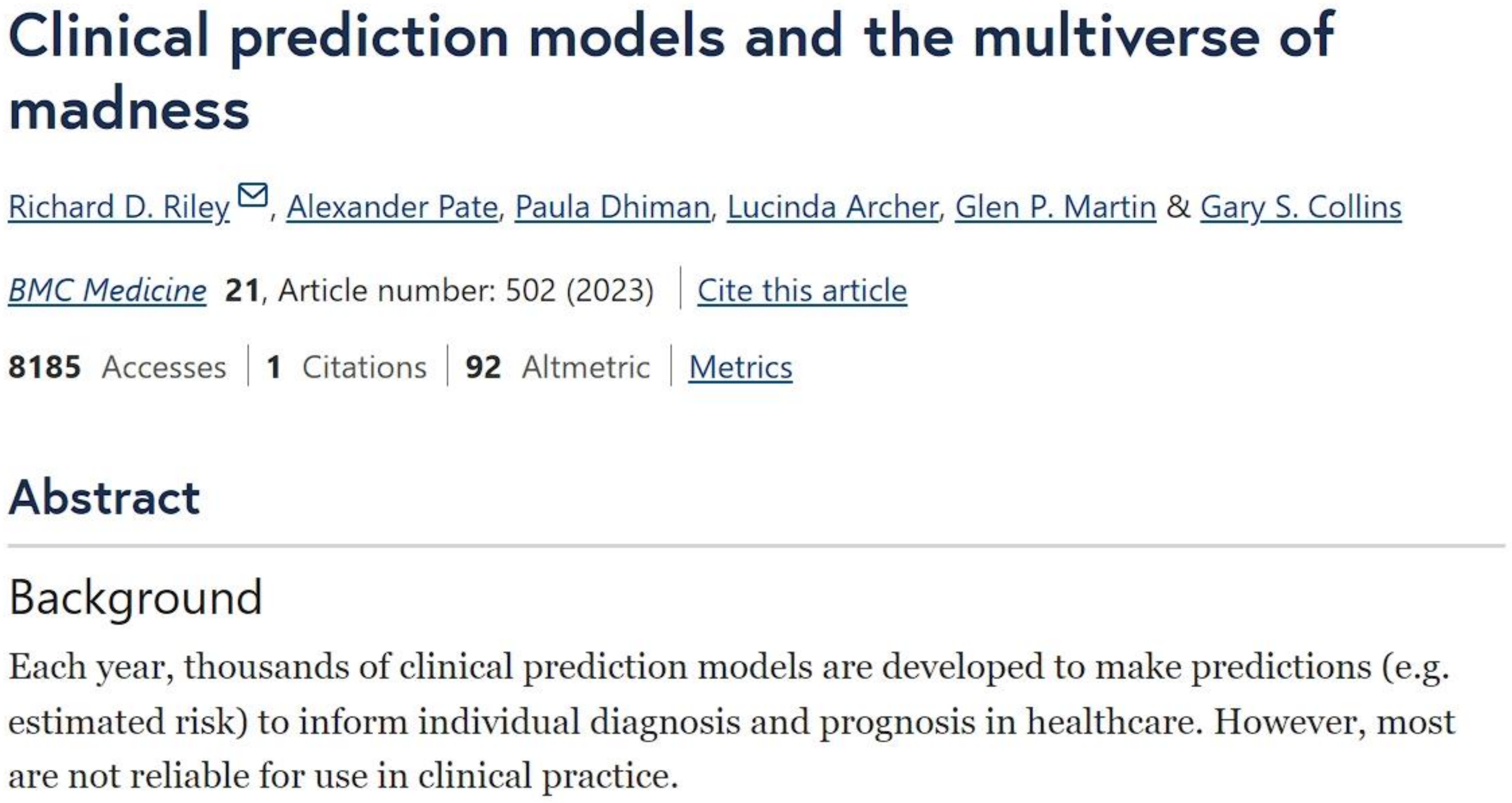
3. Hyperparameter Stability

- Hyperparameters are parameters that are set before the learning process begins.
- Stable models should have hyperparameters that are not overly sensitive to changes.
- Tuning hyperparameters should lead to gradual and predictable changes in model performance.

4. Training Stability

- The training process itself should be stable.
- This means that the model should converge to a similar solution when trained multiple times on the same dataset with the same hyperparameters.
- Unstable training processes may lead to inconsistent model performance.

Back to the presenter …



Clinical prediction models and the multiverse of madness

Richard D. Riley ✉, Alexander Pate, Paula Dhiman, Lucinda Archer, Glen P. Martin & Gary S. Collins

BMC Medicine **21**, Article number: 502 (2023) | Cite this article

8185 Accesses | **1** Citations | **92** Altmetric | Metrics

Abstract

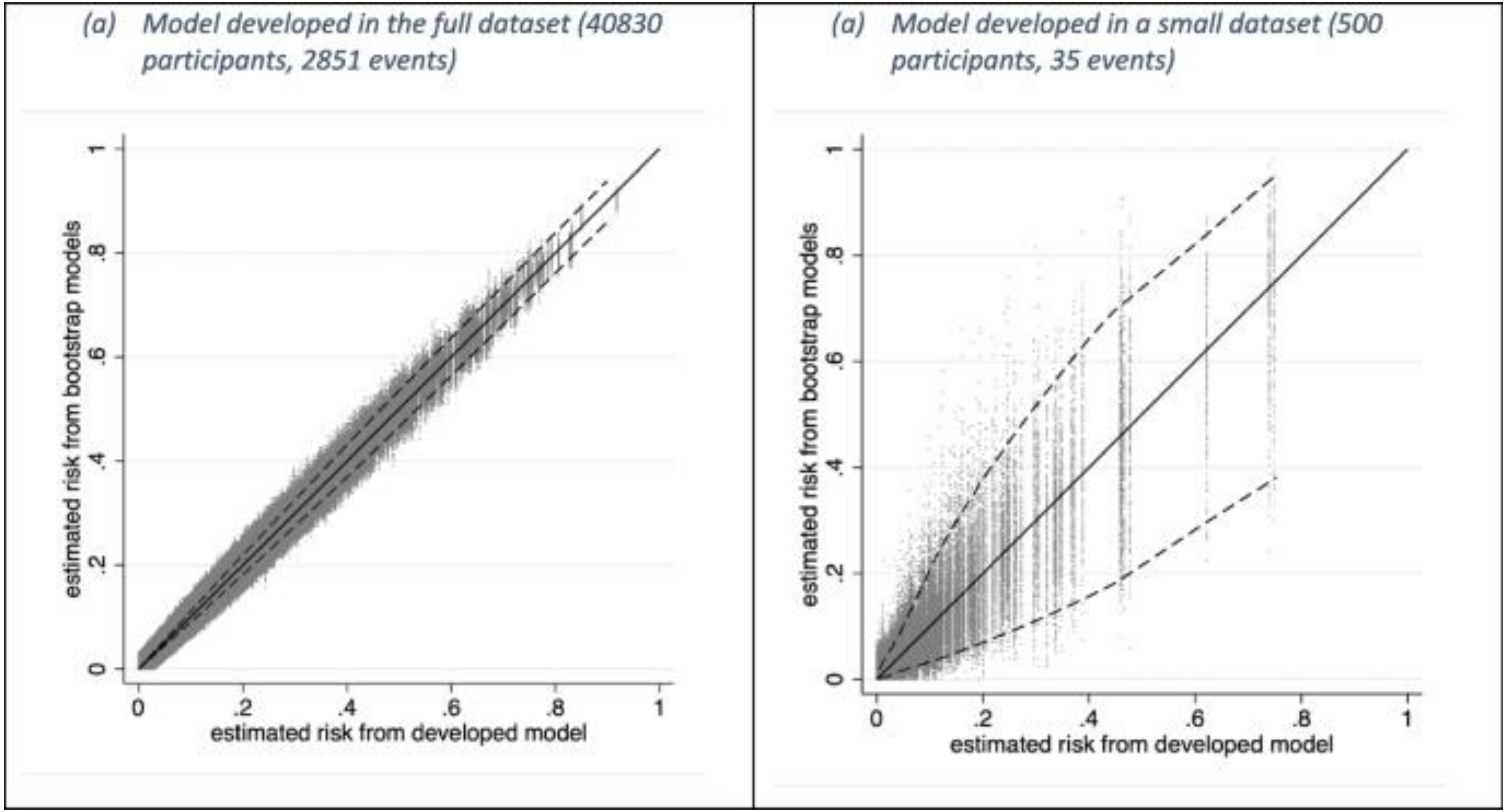
Background

Each year, thousands of clinical prediction models are developed to make predictions (e.g. estimated risk) to inform individual diagnosis and prognosis in healthcare. However, most are not reliable for use in clinical practice.

Summary of the paper

- Prediction model creation is dependent on the sample size and data used.
- Using a different sample from the same population could result in a very different model, even with the same development methods.
- Each model represents one possibility among many potential models for that sample size.
- An individual’s predicted value can vary significantly across different potential models (the multiverse).
- Different sample of the same size used from the same overarching population, the developed model might look very different (e.g. in terms of included predictors, predictor effects , regression equation, tuning parameters, tree characteristics and splits)

Same dataset; different sample size



logistic regression model (with a lasso penalty)

Suggested Solutions

- Researchers should seek to reduce model uncertainty by adhering to minimum sample size requirements that aim to precisely estimate the overall risk or mean value in the population.
- Instability can be assessed using bootstrapping and instability plots.
- Healthcare researchers should use large datasets to reduce instability.
- Large datasets are crucial for reliability across subgroups and improving model fairness in practice.

Limitations of Bootstrapping

- The bootstrap process should closely mimic the steps of prediction model development to fully capture instability.
- This includes data splitting, variable selection, missing data handling, and model tuning.
- The process may be computationally intensive, especially for deep learning methods.
- The quality of the bootstrap process depends on the representativeness of the development sample for the target population.
- Small or non-representative bootstrap samples may underestimate actual instability.
- External validation in new, sufficiently large data samples is needed for evaluations in other populations.

Limitations of this paper

- Focus has been on instability in models developed using the same approach and predictors.
- Instability discussion pertains to the developed model for one chosen target population.
- This simulates placing the original researchers in each scenario, using the same protocol and analysis plan.

- Greater diversity in the multiverse would occur if researcher preferences were considered (e.g., different modeling strategies, handling missing data, candidate predictors).
- Generalizability and transportability concern applying the model to different target populations (e.g., different settings, countries).