R software tools for the generation of synthetic data from real-world data

Michael Kammer, Medical University of Vienna

michael.kammer@meduniwien.ac.at

What is synthetic data?

Synthetic data is artificial data that is generated from original data and a model that is trained to reproduce the characteristics and structure of the original data. This means that synthetic data and original data should deliver very similar results when undergoing the same statistical analysis. The degree to which synthetic data is an accurate proxy for the original data is a measure of the utility of the method and the model.

Definition from the European Data Protection Supervisor

https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en

Why synthetic data?

Sharing data advances science – wouldn‘t that be nice!

- Increases reproducibility, transparency and trustworthiness of results
- Lets others create further insights on the (same) data (re-use)
- Combine datasources to improve generalizability
- Use for demonstration or validation purposes
- Easier to analyse than e.g. federated or centralized data

…but we can rarely do it! Synthetic data to the rescue!

Data generation is key

Sharing an original dataset is difficult

- It may be enough to share a „similar“ dataset*
- Tradeoff between fidelity, utility and privacy

75.47292	-3.181471	0.1914619
75.62670	-3.645042	0.1984420
74.97358	-2.701850	0.1816263
75.01954	-3.273654	0.1927918
75.97795	-3.802442	0.1886332

* Measuring similarity can be done in many ways:

- by reproduction accuracy of a single intended application
- by domain angostic scores for dataset

Alaa AM, et al. How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models, 2021 [arXiv:2102.08921 p.]

Data generation is key

Sharing an original dataset is difficult

- It may be enough to share a „similar“ dataset
- Tradeoff between accuracy and privacy

75.47292	-3.181471	0.1914619
75.62670	-3.645042	0.1984420
74.97358	-2.701850	0.1816263
75.01954	-3.273654	0.1927918
75.97795	-3.802442	0.1886332

Synthetic data requires a generative model

- May be based on changes of the data, or a statistical model
- Sharing the model as alternative to sharing a single dataset
- May let others adapt synthetic data to their needs

$$f$$

Use case: simulation studies

Received: 12 August 2022 | Revised: 9 December 2022 | Accepted: 22 January 2023
DOI: 10.1002/bimj.202200222

Biometrical Journal

RESEARCH ARTICLE

Phases of methodological research in biostatistics—Building the evidence base for new methods

Georg Heinze[1] | Anne-Laure Boulesteix[2] | Michael Kammer[1,3] | Tim P. Morris[4] | Ian R. White[4] | on behalf of the Simulation Panel of the STRATOS initiative

III	… comparing a relatively new method with competitors and demonstrating its use in practice; it will consider a wide range of applications.	… simulations with wide range of scenarios and different outcome types (ideally set up as neutral comparison studies), realistic comparative example data analyses.	… in which settings (among many) a method can be safely used and in which it outperforms competing methods.
IV	… summarizing the evidence about a method, also in comparison with competing methods; uncovering previously unknown behavior of the method in complex data analyses; considering an extended range of possible and actual applications.	… a review of the existing evidence about a method, simulations with extended range of scenarios, complex comparative example data analyses.	… when a method is and when it is not the preferred method; what diagnostics are available and which pitfalls may occur with its application.

Simulation studies are key for biostatistical research

Received: 12 August 2022 | Revised: 9 December 2022 | Accepted: 22 January 2023
DOI: 10.1002/bimj.202200222
Biometrical Journal

RESEARCH ARTICLE

Phases of methodological research in biostatistics—Building the evidence base for new methods

Georg Heinze[1] | Anne-Laure Boulesteix[2] | Michael Kammer[1,3] | Tim P. Morris[4] | Ian R. White[4] | on behalf of the Simulation Panel of the STRATOS initiative

Received: 29 November 2017 | Revised: 23 August 2018 | Accepted: 2 November 2018
DOI: 10.1002/sim.8086
WILEY Statistics in Medicine

TUTORIAL IN BIOSTATISTICS

Using simulation studies to evaluate statistical methods

Tim P. Morris[1] | Ian R. White[1] | Michael J. Crowther[2]

OPEN ACCESS Freely available online
PLOS ONE

A Plea for Neutral Comparison Studies in Computational Sciences

Anne-Laure Boulesteix[1*¶], Sabine Lauer[1¶], Manuel J.A. Eugster[2,3]

1 Department of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-University of Munich, Munich, Germany, **2** Department of Statistics, Ludwig-Maximilians-University of Munich, Munich, Germany, **3** Helsinki Institute for Information Technology, Department of Information and Computer Science, Aalto University, Espoo, Finland

How to get data for simulation studies?

Real study data

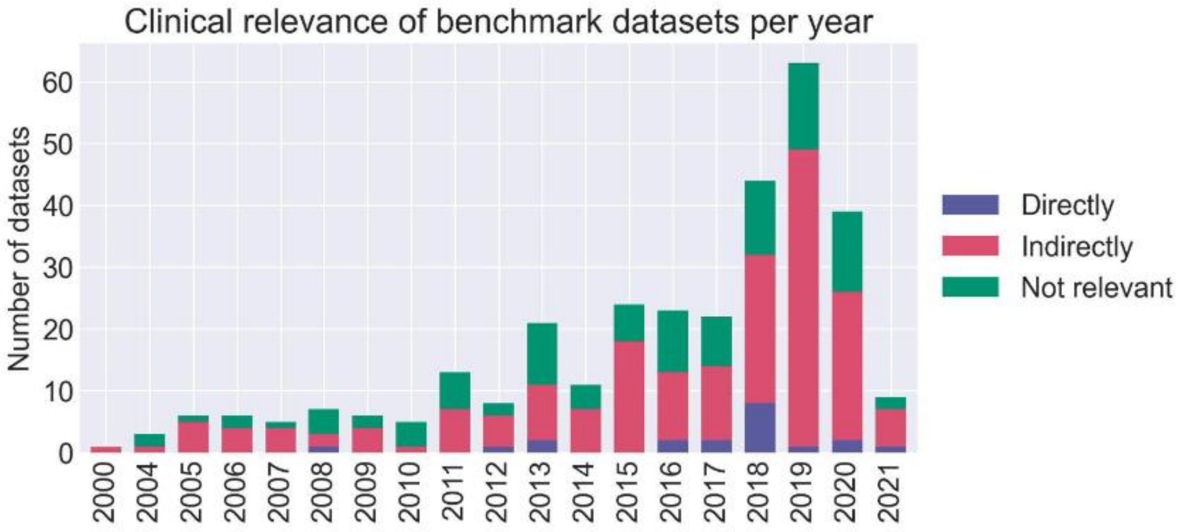
- Rarely shared, usually an example

Benchmark datasets

- „Many“ available…
- …but little control over properties

Data generators

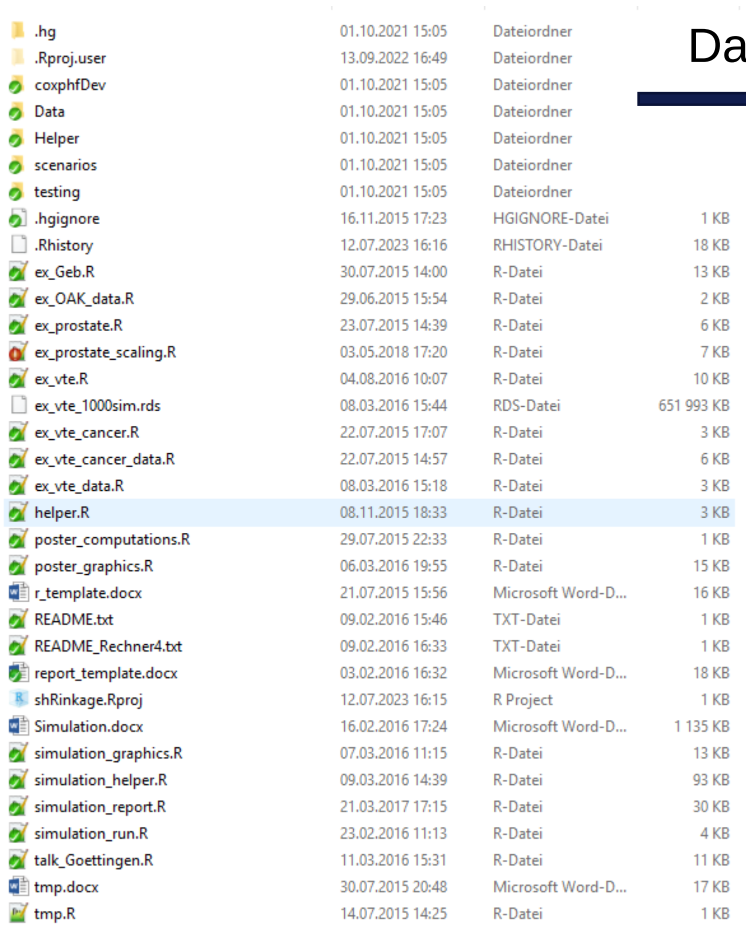
- Allow tweaking of properties…
- …but difficult to create, often ad-hoc



Blagec K, et al. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. J Biomed Inform. 2023;137:104274.

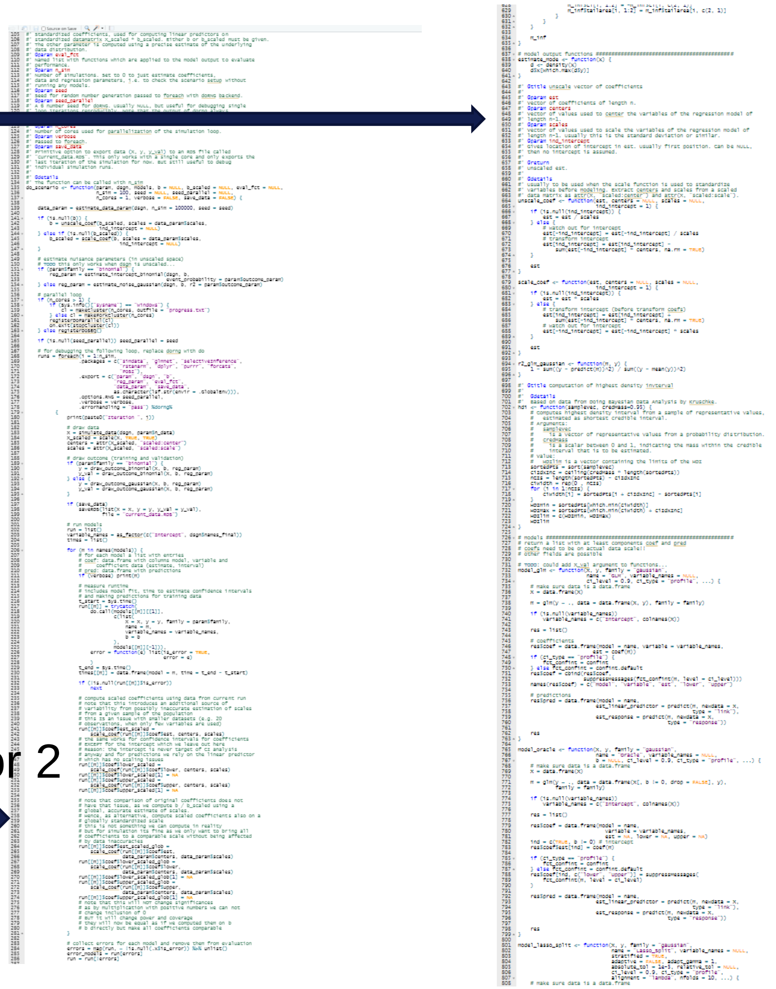
Data generation in practice: we have all been there

Many files



Data generator 1

Data generator 2



Can we improve?

Yes, by sharing purposeful data generators

- Avoid ad-hoc mistakes by sharing code based on „lessons learnt“
- Makes writing of simulation code easier
- Standardization increases familiarity

```
{r}
library(simulation_library)

design = simulation_design(...)
data = simulate_data(design)
resuts = analyse(data)
```

Can we improve?

Yes, by sharing purposeful data generators

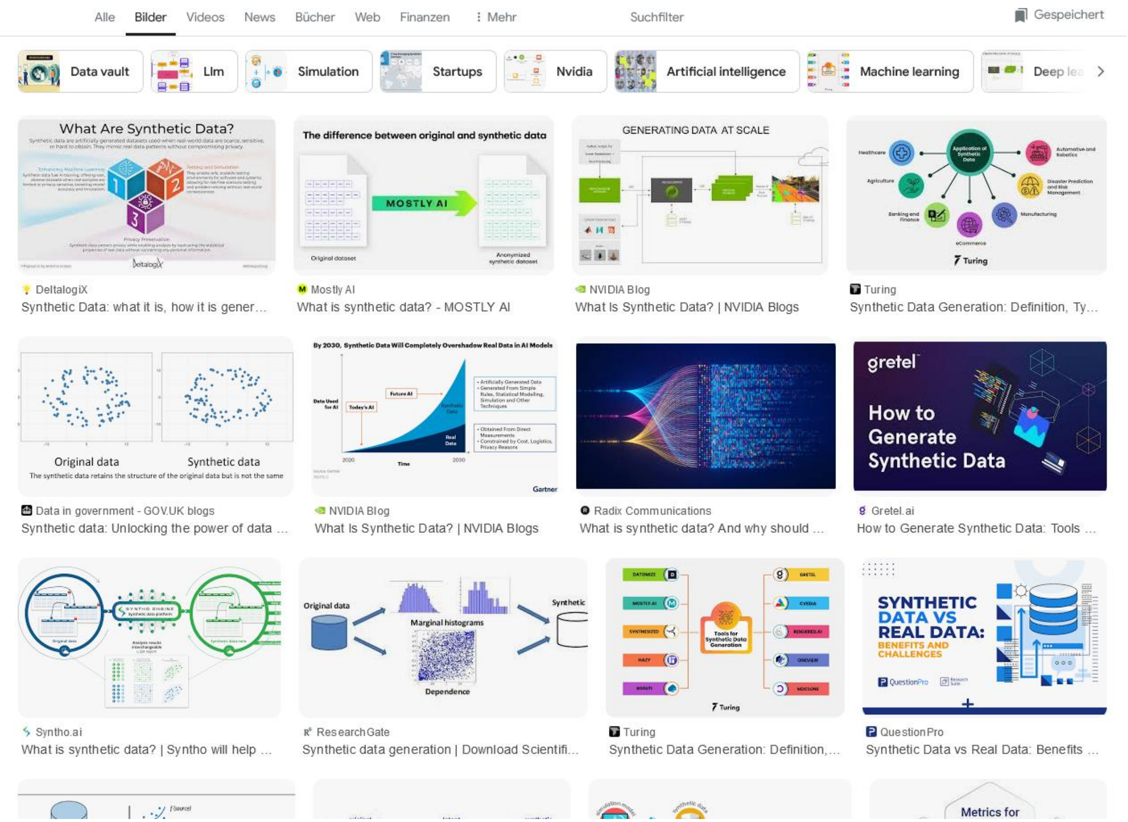
- Avoid ad-hoc mistakes by sharing code based on „lessons learnt“
- Makes writing of simulation code easier
- Standardization increases familiarity

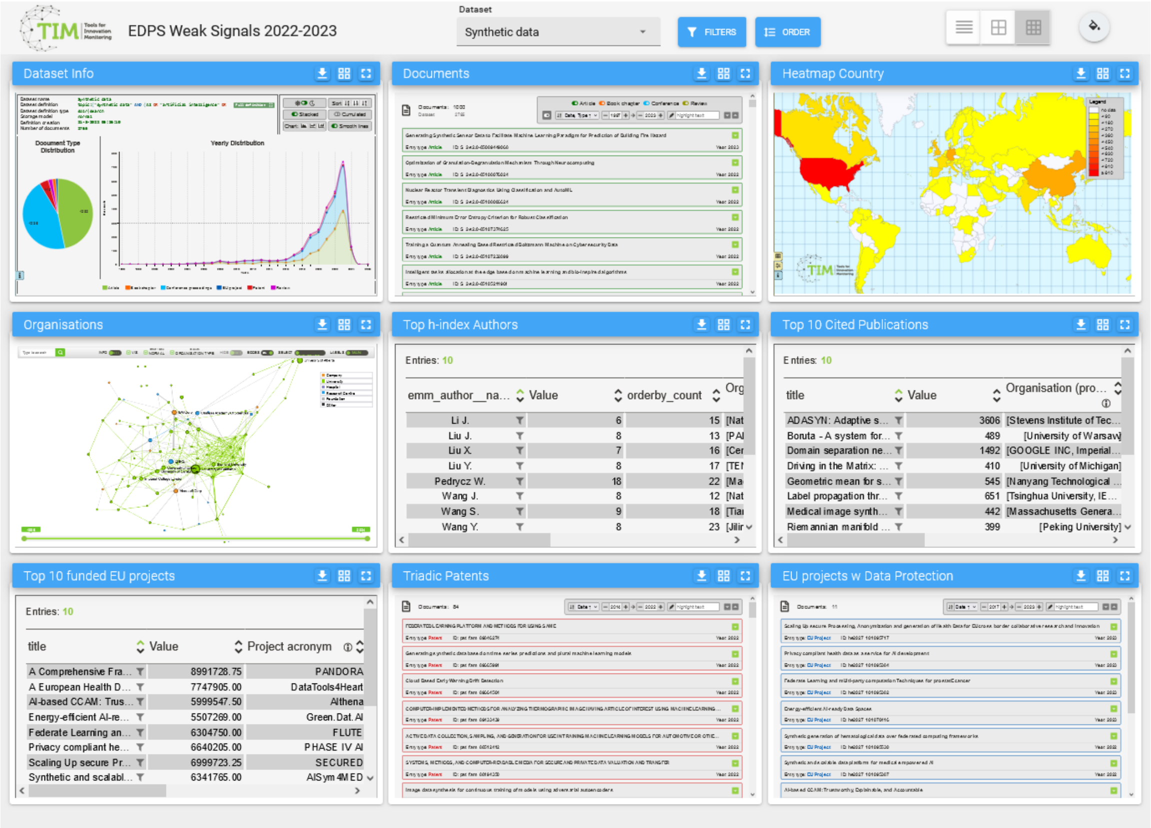
But…

- Avoid oversimplification and monopolization
- Dissemination
- Creating and documenting reliable code takes a lot of time!

A first step: what is out there?

Synthetic data is a hot topic





Google search „Synthetic data“ on 8.1.2025

https://www.timanalytics.eu/TimTechPublic/dashboard/index.jsp#/space/s_1836?ds=224919

A first step: what is out there?

Scoping review of R packages to support data generation

- Focus on packages that support data generation for…

  …Tabular data common in biostatistical research / simulations

  …Multiple, correlated variables of different types

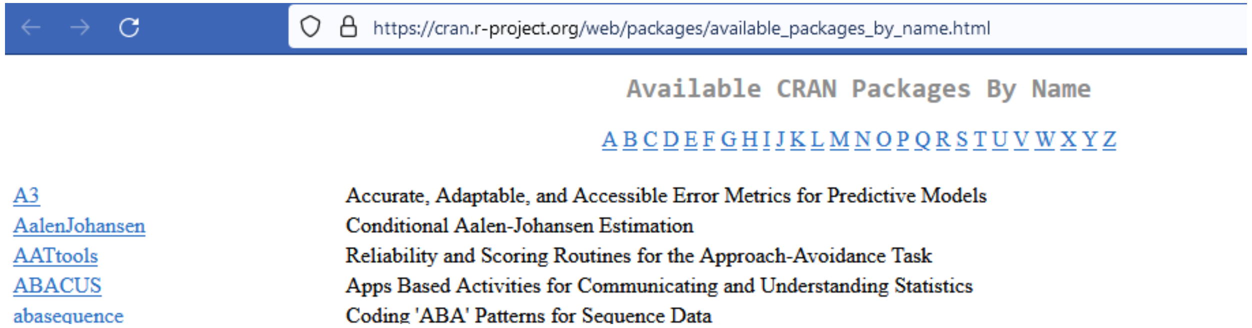
  …No timeseries, counterfactuals, images, text, physical processes, …

Declaration of Biases:

- **I authored the *simdata* package.**
- **I am not an expert on many of the methods.**

Search strategy

- Search packages on CRAN
  - Via static html page
  - Via *pkgsearch*
  - Manual finds
- Keywords:
  - *simulat** (includes terms such as simulation, simulate, simulating…)
  - *generat**
  - *synth**
- Search date: 18.07.2023 with subsequent updates



Search and screening results

Search

1193 found

43 removed from CRAN

1150 screened

Screening stage 1

632 false positives
290 out of scope
44 model simulation
…

136 for Stage 2

Screening stage 2

94 low priority
12 med priority
6 special
10 exclude

14 top priority

3 additions later on

Selected packages (full list at https://osf.io/9gfns)

	Type	Functionality	Maintenance	Documentation
faux	Framework	High	High	High
SimDesign	Framework	High	High	High
simPop	Framework	High	High	Med
simstudy	Framework	High	High	High
synthpop	Framework	High	High	High
fabricatr	Framework*	Med	High	High
simFrame	Framework*	High	High	High
simpr	Framework*	High	High	High
simulator	Framework*	High	High	High
bigsimr	Generator	High	Med	Low
covsim	Generator	High	Med	Med
PoisBinOrdNonNor	Generator	High	Med	Med
SimCorrMix	Generator	High	Low	High
SimJoint	Generator	High	Med	Med
SemiArtificial	Generator	Med	Med	Med
Modgo	Generator	High	High	High
arf	Generator	High	High	Med
simdata	Framework	High	High	High

* no data generation

Data generation methodologies

Power polynomials

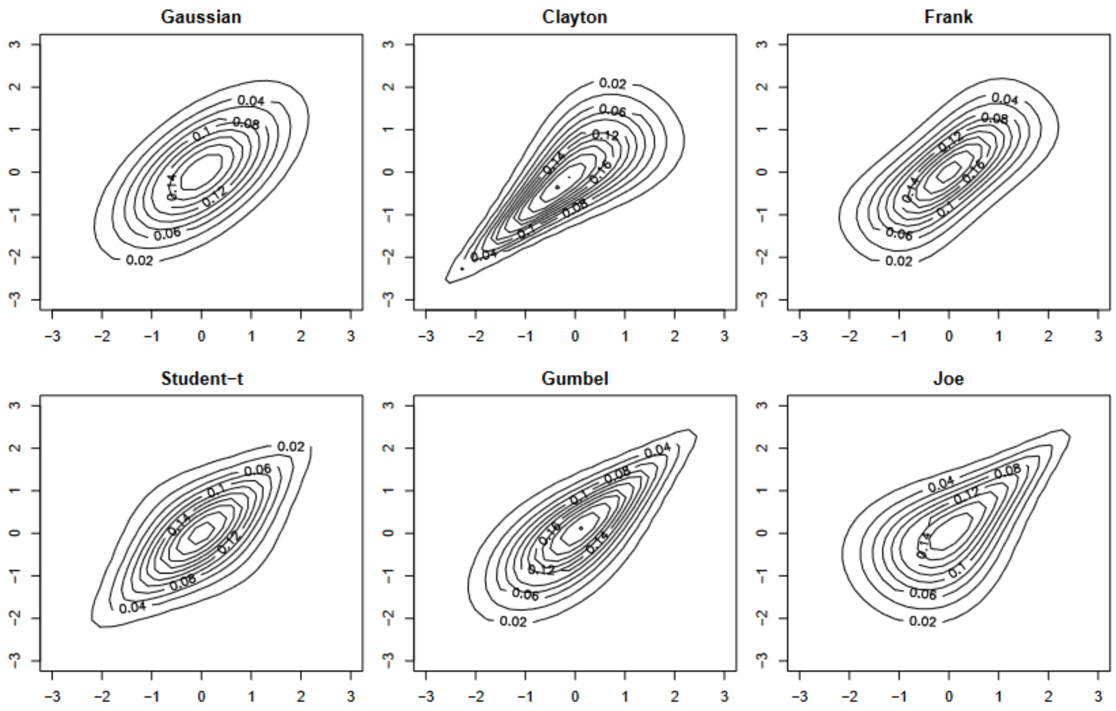
- Specify non-normal continuous data as $X = a + bZ + cZ^2 + dZ^3$, $Z \sim N(0,1)$ (Fleishman 1978)
- Extended to multivariate data (Vale 1983)
- Extended to further moments (Headrick 2002)
- Long standing in social science and psychometrics (Headrick 2010)
- Implementations in *PoisBinOrdNonNor, SimCorrMix, SimDesign*

Fleishman AI. A method for simulating non-normal distributions. Psychometrika. 1978;43(4):521-32.
Headrick TC. Statistical Simulation - Power Method Polynomials and Other Transformations: Chapman & Hall; 2010.
Headrick TC. Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. Computational Statistics & Data Analysis. 2002;40(4):685-711.
Vale CD, Maurelli VA. Simulating multivariate nonnormal distributions. Psychometrika. 1983;48(3):465-71.

Data generation methodologies

Copula based

- Joint multivariate distribution with marginals $Uniform[0,1]$
- Many applications in diverse fields
- NORTA implemented in *covsim, faux, bigsimr, simdata, modgo* (also similar)
- Other approaches in *covsim, simstudy*



Stenšin (2022)

Stenšin A, Bloznelis D. Copulas and Portfolios in the Electric Vehicle Sector. Journal of Risk and Financial Management. 2022;15(3).

Data generation methodologies

Normal to anything (NORTA, Cario 1997)

- Given distribution functions $F_i(s) = P(X_i \leq s)$ and correlation matrix $\Sigma_{\boldsymbol{X}}$:

  Generate multivariate standard normal vectors $\boldsymbol{Z}$ with correlation $\Sigma_{\boldsymbol{Z}}$

  Simulate $\boldsymbol{X}'$ via $X_i' = F_i^{-1}\left(\Phi(Z_i)\right)$, where Φ is the distribution function of the standard normal distribution

- To achieve desired final correlation matrix, $\Sigma_{\boldsymbol{Z}}$ must be chosen appropriately

  Solve univariable optimisation problem for each pair of variables

  Ensure result is a positive-definite correlation matrix

- *modgo* uses a polychoric correlation for binary / categorical data

Cario MC, Nelson BL. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Evanston, Illinois: Northwestern University, Sciences DoIEaM; 1997.

Data generation methodologies

Model based (conditional specification)

- Approximate joint distribution by sequence of conditional distributions (models)
    - Start by synthetizing x_1' from x_1 (i.e. sample with replacement)
    - Fit model for x_2 using observed x_1 and sample from fitted model to obtain x_2'
    - Continue for all variables
- Could include X_{open} only used for synthesis
- Used in *synthpop* (Nowok 2016), with default synthetizer model CART

Nowok B, Raab GM, Dibben C. synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software. 2016;74(11).

Data generation methodologies

Model based (Machine learning based)

- Data generation based on RBF networks and random forests
- Used in *SemiArtificial, arf*

Other approaches not covered in scoping review and not in R?

- Adverserial networks
- Autoencoders

Data generation methodologies

	Power polynomials	*NORTA*	*Model based*
Marginals	First 4 moments	Full marginals	Not directly
Correlation	Yes	Yes	Not directly
Relations	Linear	Linear	Non-linear
Categories	Categorization	Quantiles	Modeling
Input	Moments + Correlations	Quantiles + Correlations	Full dataset
Setup	One-time	One-time	Repeated / One-time

Artificial data comparison: setup

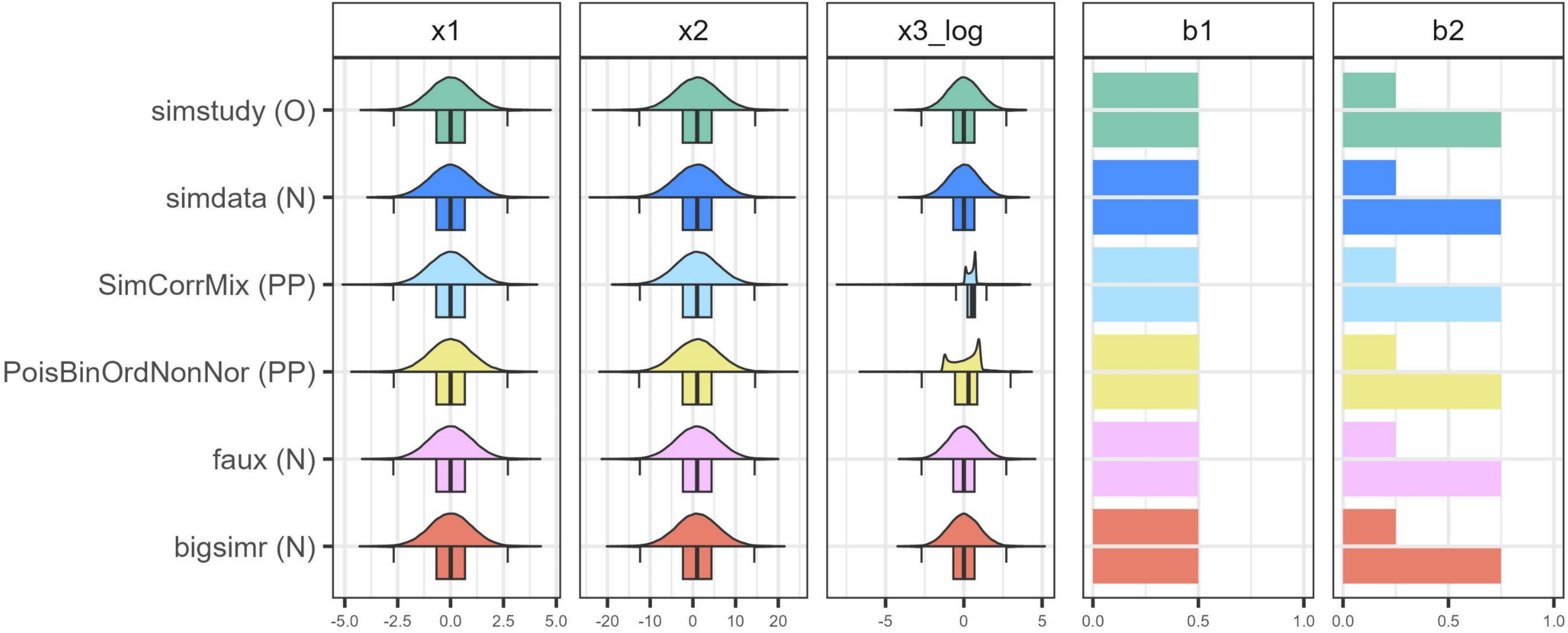
Here: low-dimensional „realistic“ data

- 5 variables: $N(0,1), N(1,5), LogN(0,1), Bern(0.5), Bern(0.25),$
- Target pairwise Pearson correlation: 0.2
- 100,000 observations

Notes

- *Synthpop, SemiArtificial, modgo, arf* require the input of an original dataset and are not included for now
- For all packages code was quite simple to write

Artificial data comparison: marginals

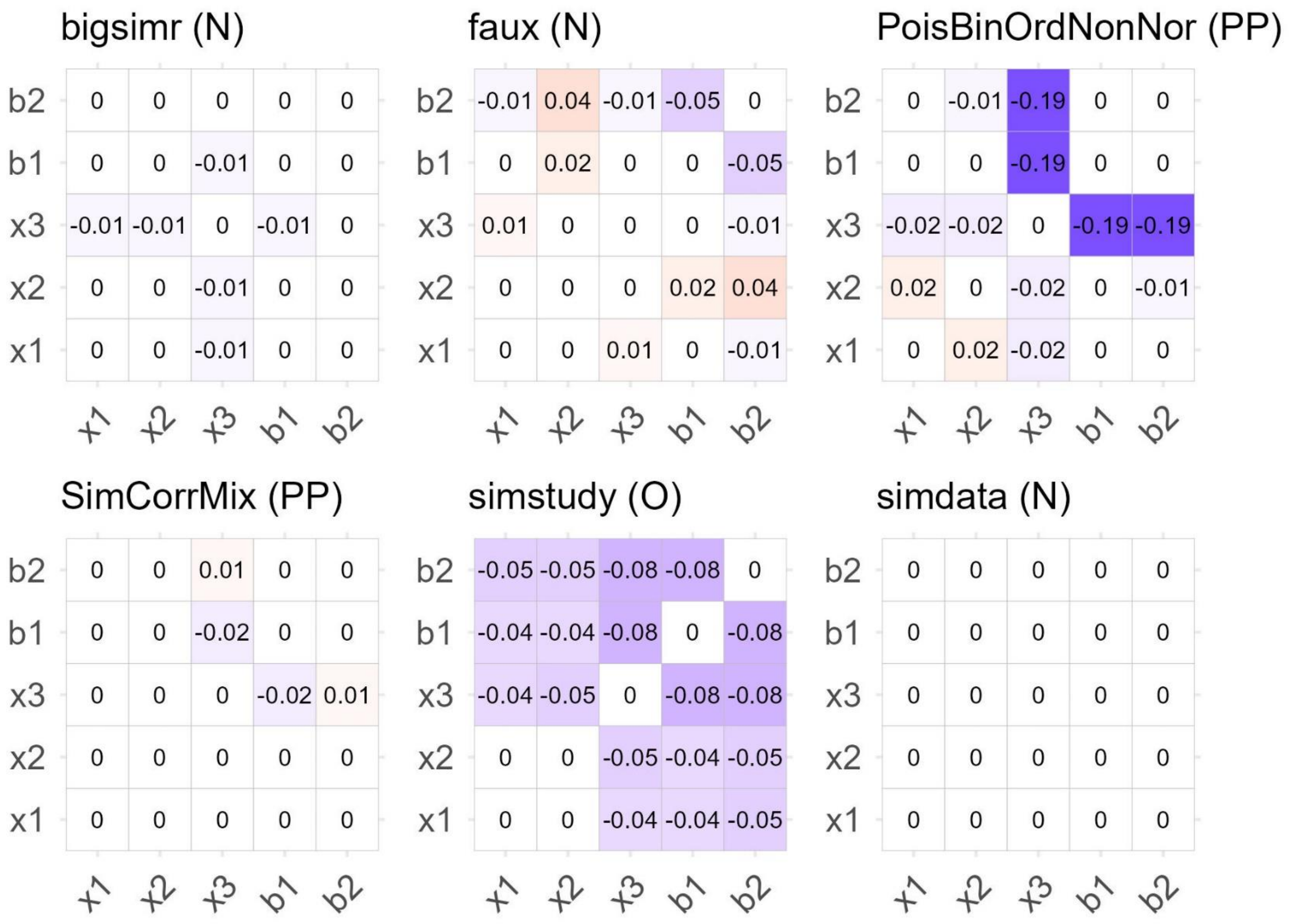


PP: Power polynomials, N: NORTA, MB: model based, O: other

Artificial data comparison: correlations

Difference to target

(achieved – 0.2)



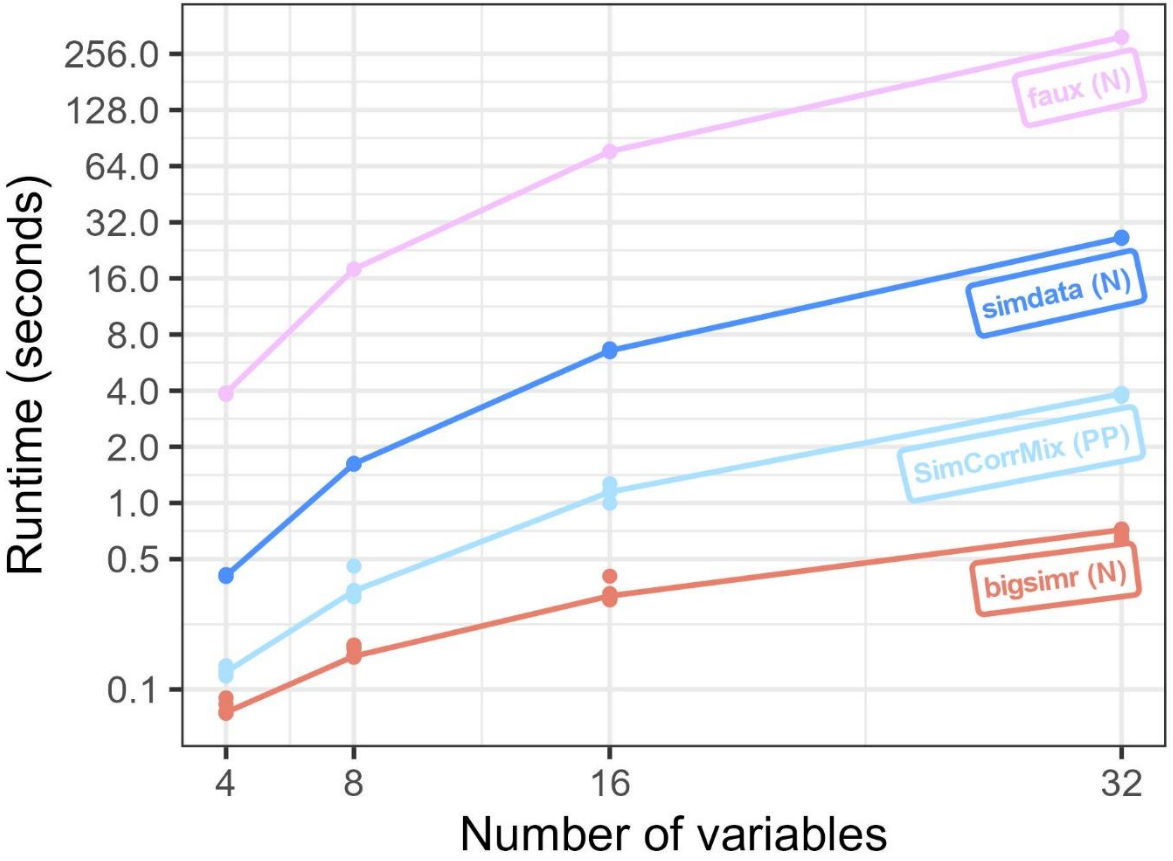
Artificial data comparison: runtime scaling

Simple data generation

- p' $N(0, 1)$ and p' $Bern(0.5)$ variables
- Target Pearson correlation: 0.2
- 100,000 observations, 10 runs

Notes

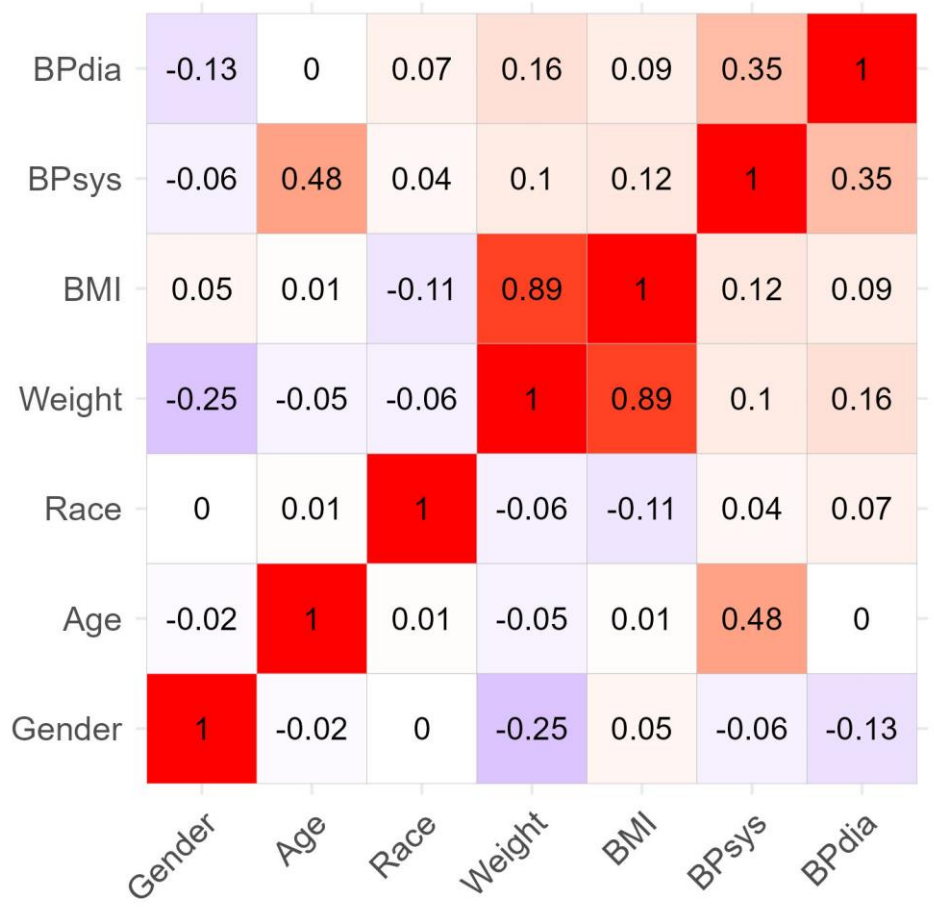
- Mostly quadratic in $p = 2p'$
- *PoisBinOrdNonNor, simstudy* excluded



Real world data: NHANES

Demograhics dataset (via *nhanesA* package)

- 4709 observations
  - 9254 initial
  - 6230 complete
  - 4709 age > 18
- 7 variables of different types
  - Gender binary
  - Race categorical (5 categories)
  - Others continuous
- Generate 100,000 samples
- Also include *SemiArtificial, synthpop, modgo, arf*



Real world data: coding notes

Specification of data generators quite different

bigsimr requires parametric description of marginals and correlations

simdata offers automated estimation of marginals, requires correlations

SimCorrMix requires estimation of moments and correlations

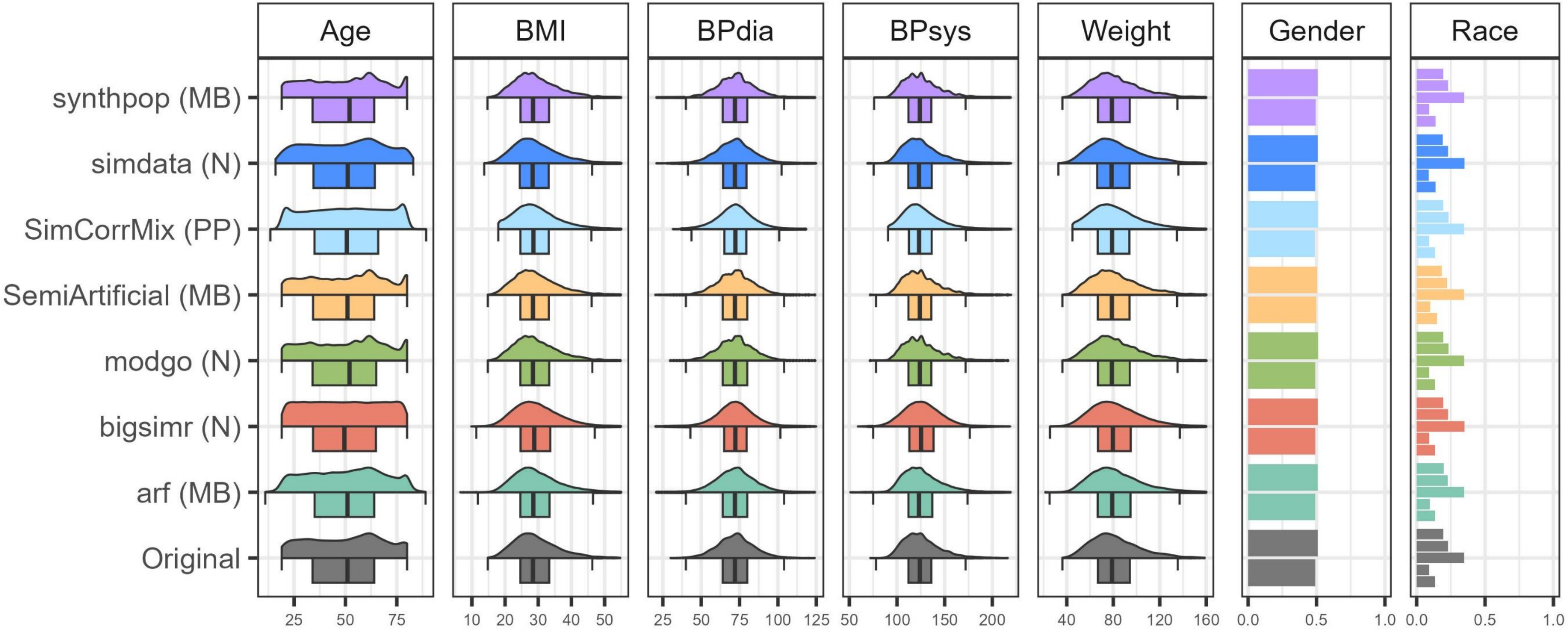
synthpop uses CART by default for each variable, only requires data

SemiArtifical only requires data, no parameters except number of trees

modgo only requires data and some simple parameters

arf only requires data and some random forest parameters

Real world data: marginals

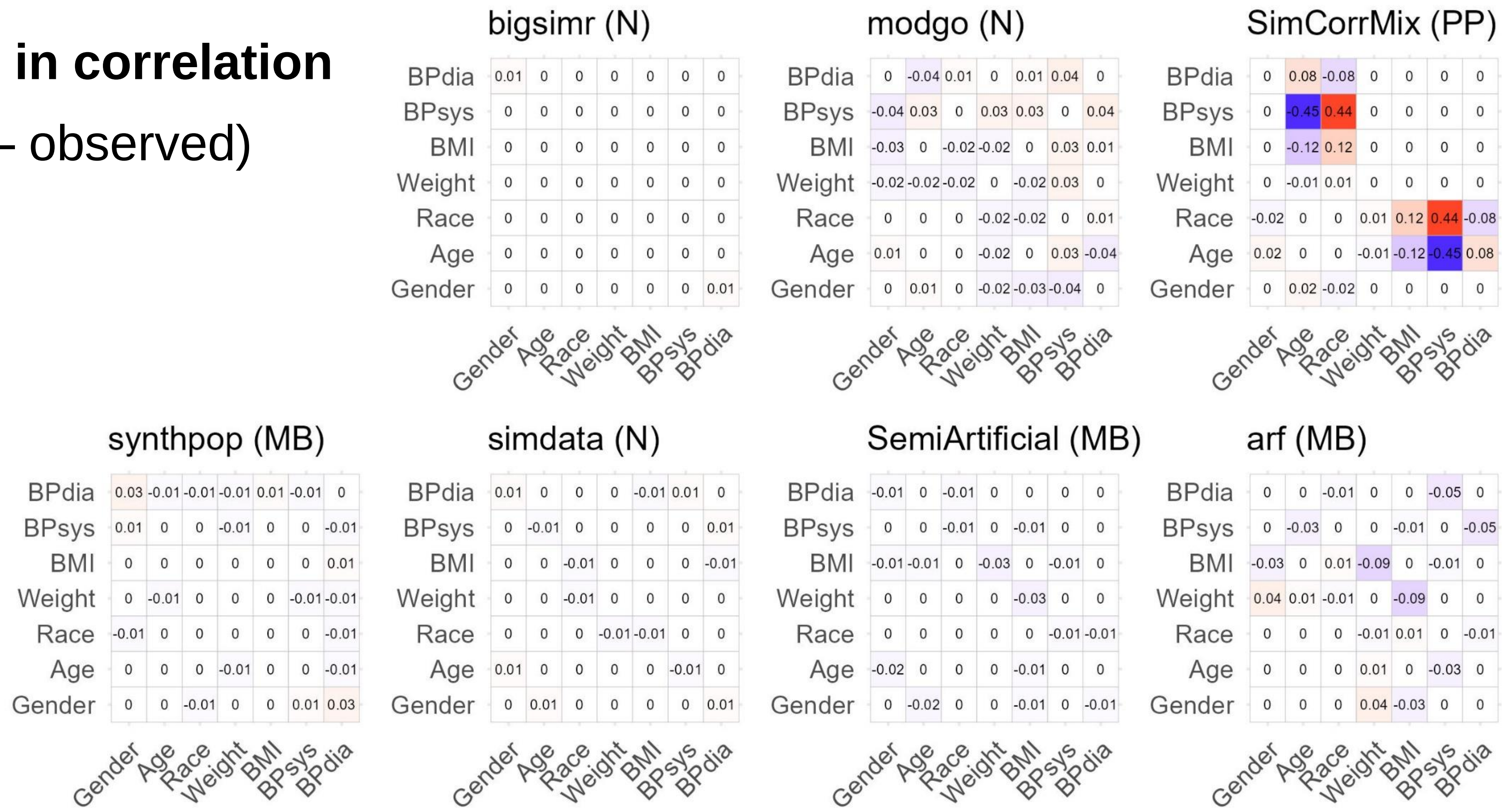


PP: Power polynomials, N: NORTA, MB: model based, O: other

Real world data: correlations

Difference in correlation

(achieved – observed)



Real world data: runtimes

Large differences in runtimes (seconds)!

	Setup only	*Setup + Generation*
bigsimr	0.006	0.113
simdata	0.679	0.895
SimCorrMix	-	15.308
synthpop	-	5.020
SemiArtificial	22.406	125.600
modgo	-	0.553
arf	19.928	21.614

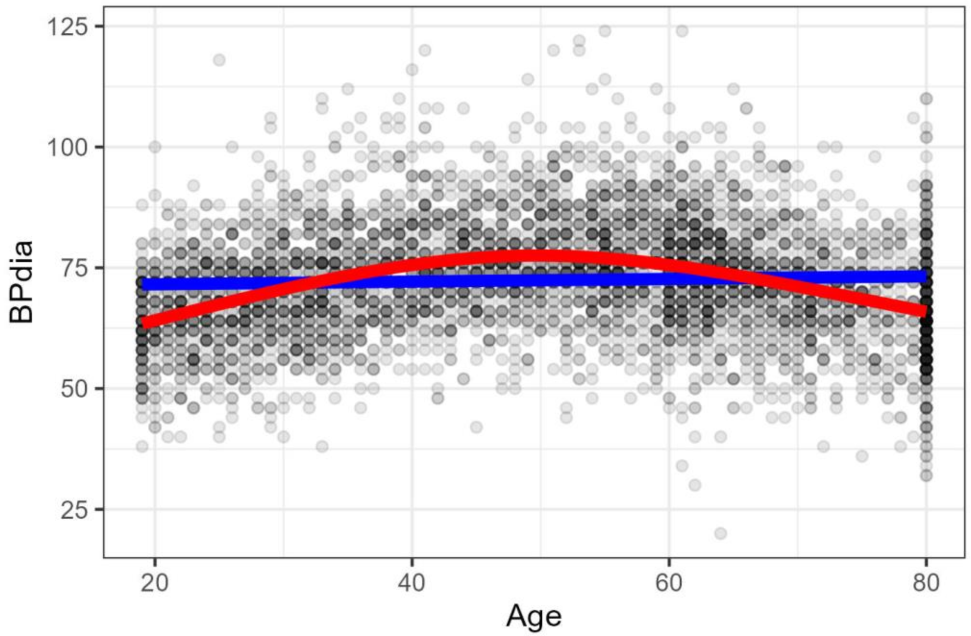
All results in seconds, medians of 10 runs, Windows 10, Intel i7-10700K@3.8Ghz

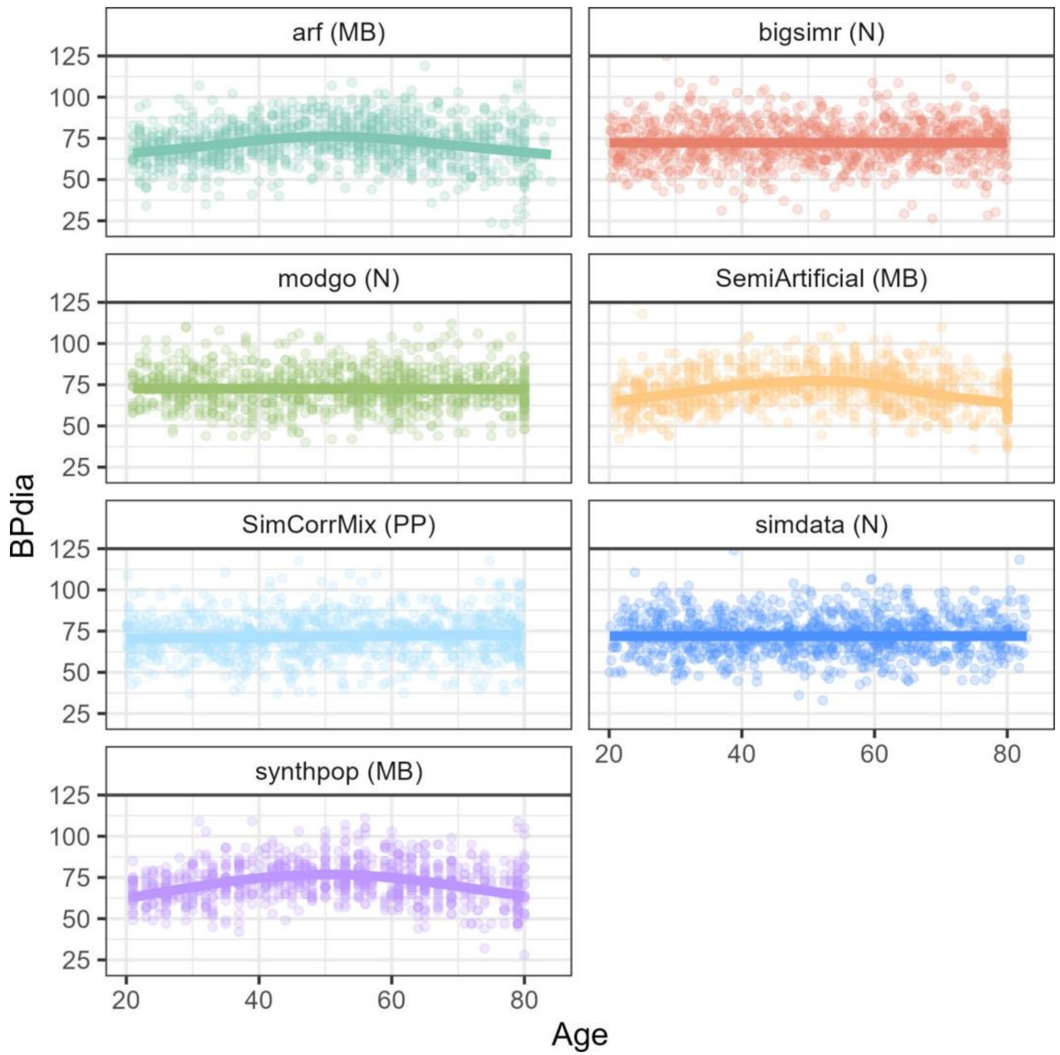
Real world data: non-linearities

Correlations imply form

Methods based on correlations

restrict form of relations





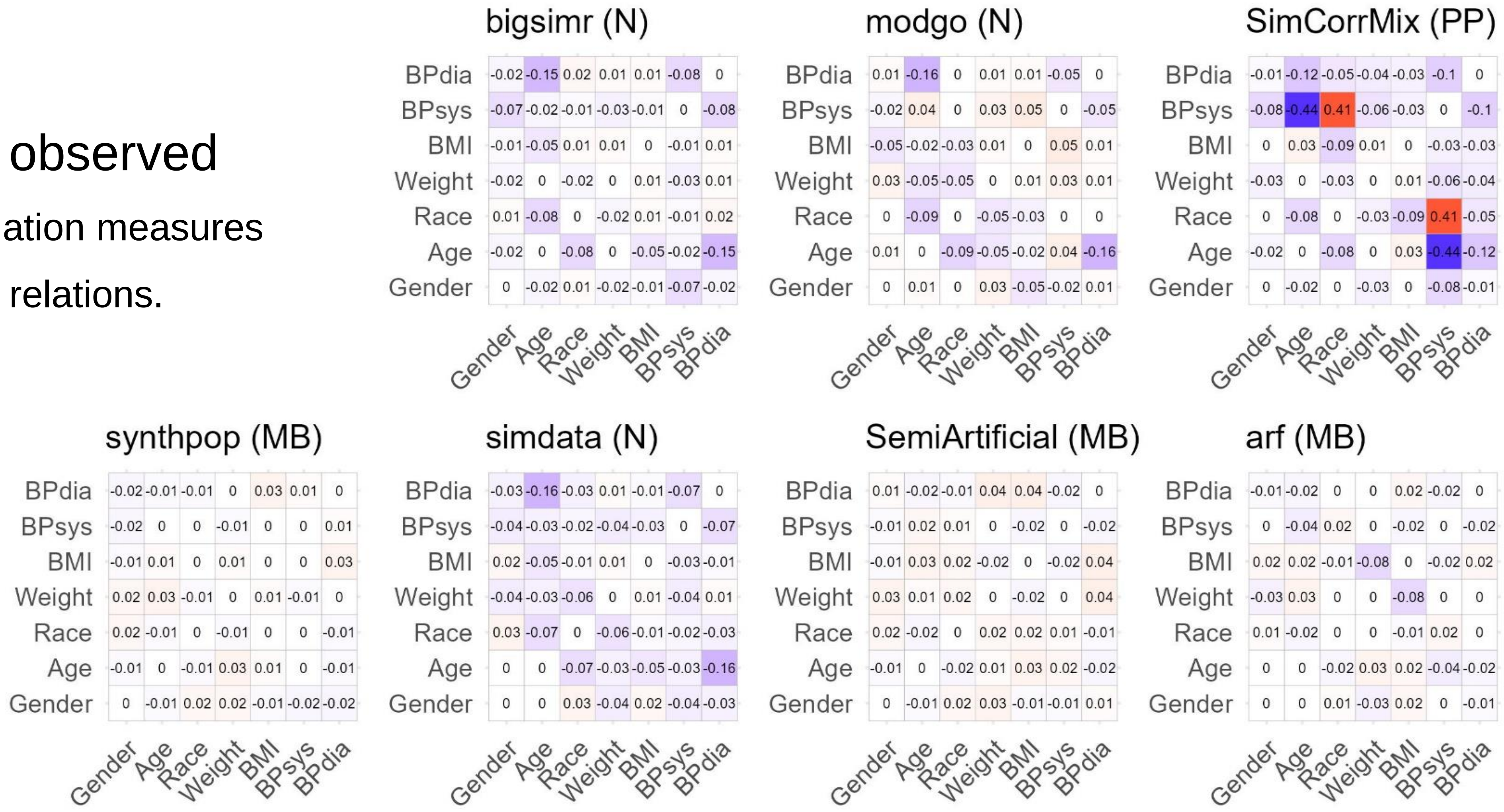
Real world data: distance correlations

Difference

achieved – observed

Distance correlation measures

also non-linear relations.



Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. The Annals of Statistics. 2007;35(6):2769–94.

Recreating regression analyses

Model 1: simple

- Outcome mean arterial pressure

  MAP = BPdia + 1/3 (BPsys – BPdia)

- Covariates Age, Gender, BMI
- **Addon**: Covariate Race

Recreating regression analyses

Model 1: simple

- Outcome mean arterial pressure

  MAP = BPdia + 1/3 (BPsys – BPdia)

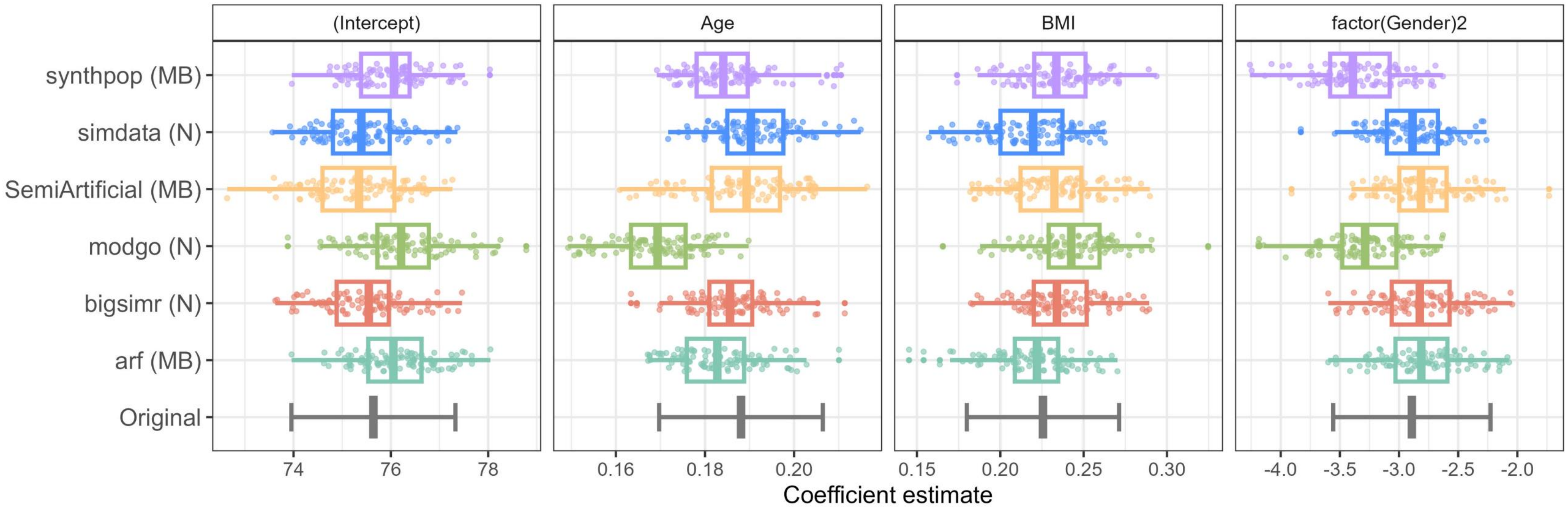
- Covariates Age, Gender, BMI
- **Addon**: Covariate Race

Model 2: spline for age

- Outcome diastolic blood pressure
- Covariate Age (3-knot restricted cubic spline)

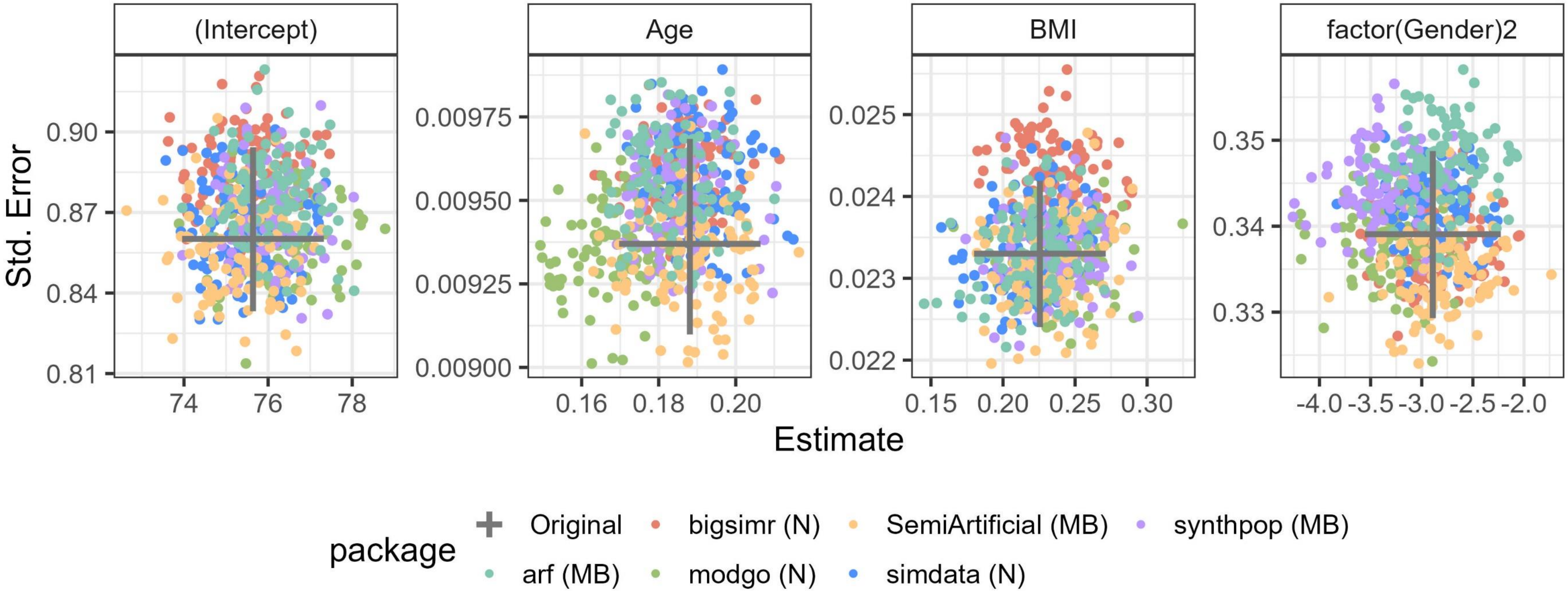
Both models re-created 100 times.

Recreating regression analyses: Model 1

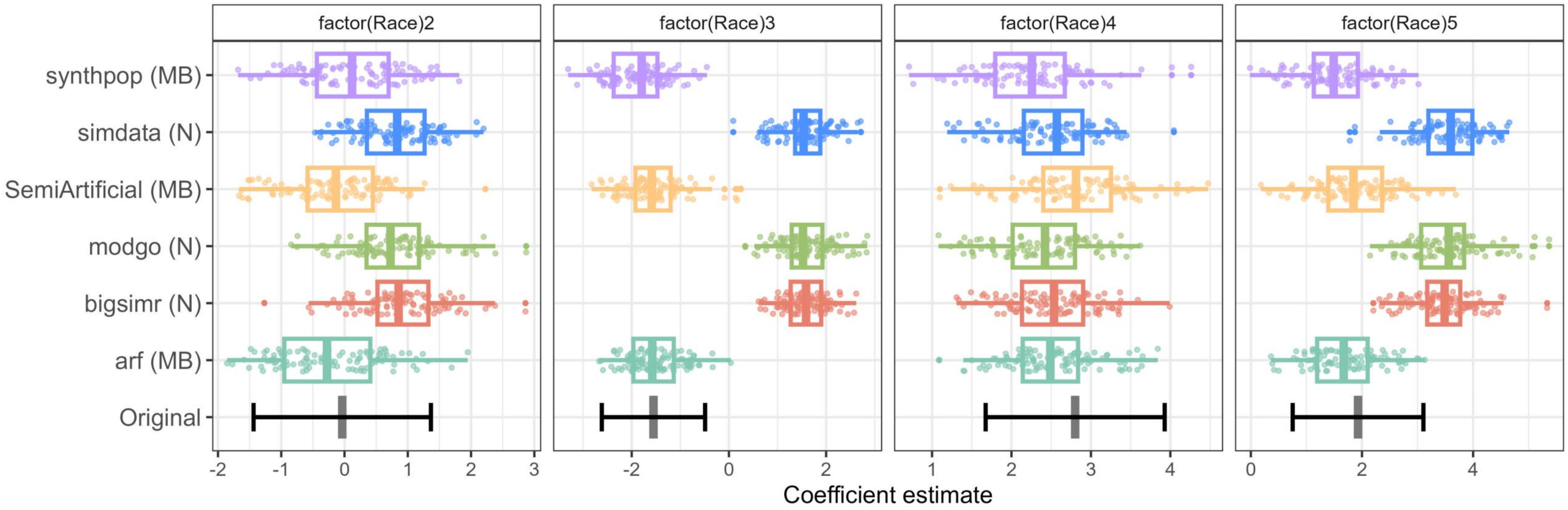


	original	bigsimr (N)	modgo (N)	SemiArtificial (MB)	simdata (N)	synthpop (MB)	arf (MB)
R^2	0.11 (0.11, 0.11)	0.11 (0.10, 0.11)	0.10 (0.10, 0.11)	0.11 (0.11, 0.12)	0.11 (0.10, 0.11)	0.11 (0.11, 0.11)	0.10 (0.09, 0.11)

Recreating regression analyses: Model 1

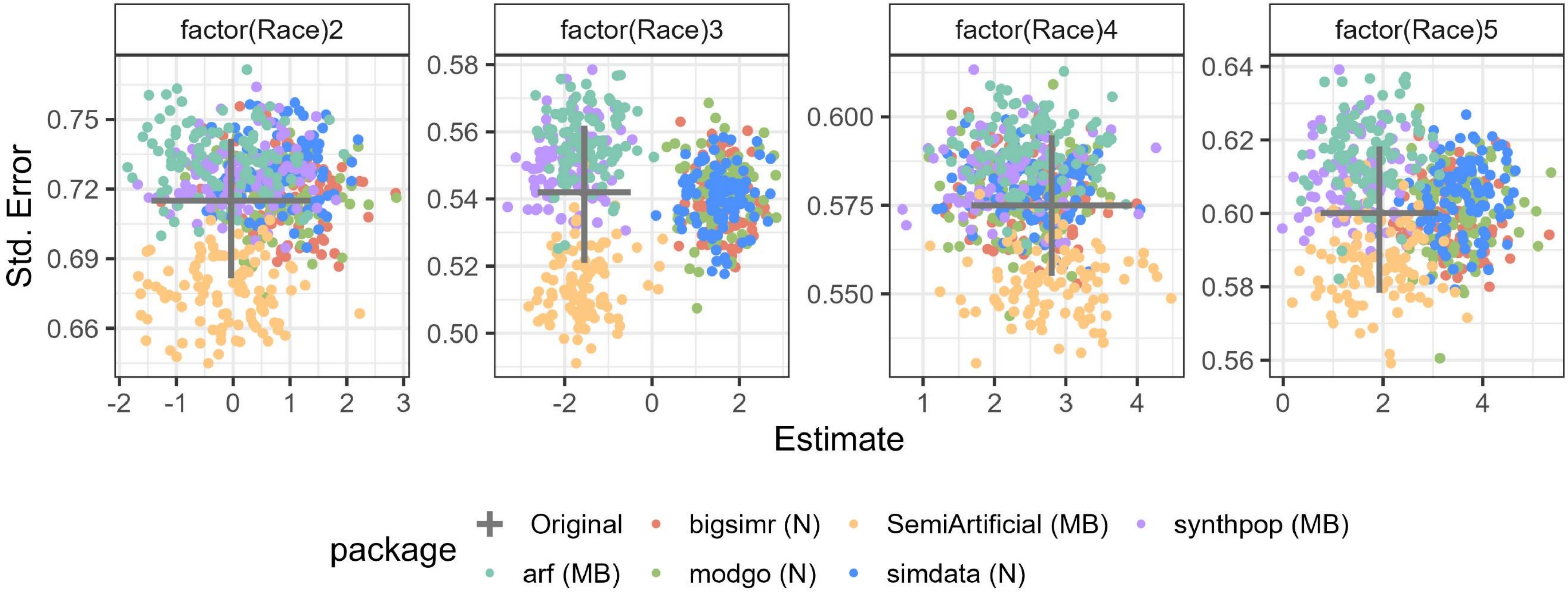


Recreating regression analyses: Model 1 addon



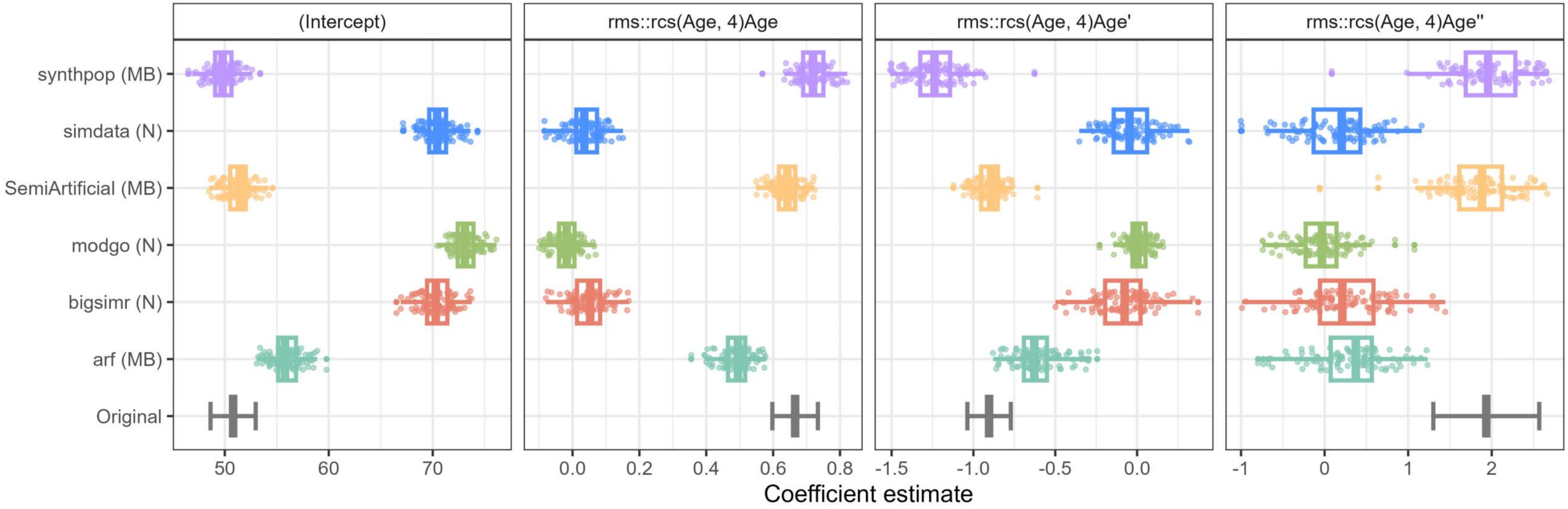
	original	bigsimr (N)	modgo (N)	SemiArtificial (MB)	simdata (N)	synthpop (MB)	arf (MB)
R^2	0.11 (0.11, 0.11)	0.11 (0.10, 0.11)	0.10 (0.10, 0.11)	0.11 (0.11, 0.12)	0.11 (0.10, 0.11)	0.11 (0.11, 0.11)	0.10 (0.09, 0.11)

Recreating regression analyses: Model 1 addon



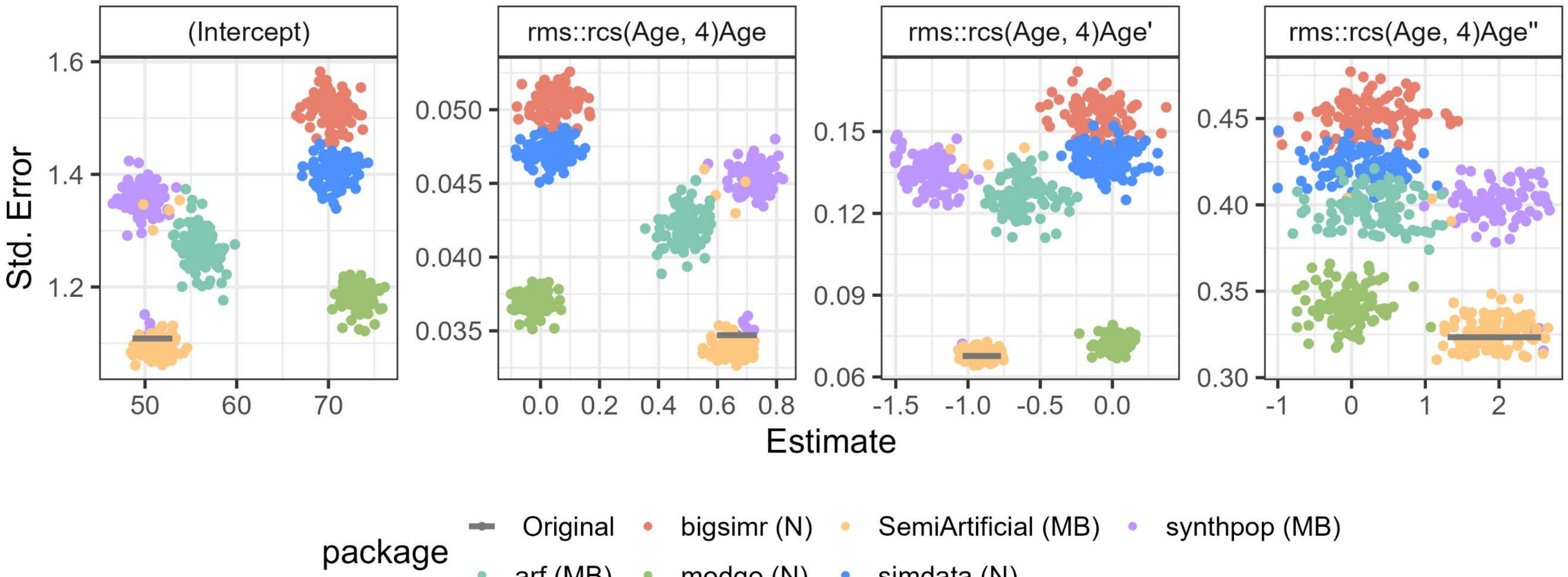
Need to model dummy variables instead for some packages? Not easy…?

Recreating regression analyses: Model 2



	original	bigsimr (N)	modgo (N)	SemiArtificial (MB)	simdata (N)	synthpop (MB)	arf (MB)
R^2	0.13 (0.13, 0.13)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.12 (0.12, 0.13)	0.00 (0.00, 0.00)	0.12 (0.11, 0.12)	0.09 (0.08, 0.09)

Recreating regression analyses: Model 2



For inference: model based depends on model…

Conclusions (selected packages)

	bigsimr	*simdata*	*synthpop*	*SemiArtificial*	*modgo*	*arf*
Goal	Artificial data	Artificial data	Faithful re-creation	Faithful re-creation	Faithful re-creation?	Faithful re-creation
Method	NORTA	NORTA	Model based	Model based	Transformation	Model based
Control	Correlations, marginals	Correlations, marginals	Relationships	Relationships	Correlations, marginals	Relationships
Ease of use	Requires Julia	Simple, but flexible	Simple, but flexible	Very simple	Simple, but flexible	Simple, but flexible
Sharing	Data generator	Data generator	Data generator	Data generator	Datasets	Data generator
Storage	Small	Small	Can be large	Can be large	-	Can be large
Setup	One-time	One-time	Each generation	One-time	Each generation	One-time
Runtime	Fast	OK	OK-Slow	Slow	OK	Slow
Privacy	OK	? (ECDF)	OK	? (RF)	? (ECDF)	?

Conclusions (selected packages)

Your requirements:

- Really fast, don‘t care about going outside R: *bigsimr*
- Marginals + correlation, fast, no actual data, special functionality: *modgo*
- Marginals + correlation, fast, no actual data, share generator: *simdata*
- Faithful recreation, flexibility, have data, popular, privacy: *synthpop*
- Faithful recreation, have data, don‘t care about speed: *SemiArtificial* / *arf*

Discussion

No single best data-generating mechanism

- Different packages address different aims
  - Re-create existing analysis or fully generative model?
  - Non-linear relations require model based approaches
- All promising packages offer good functionality
  - But runtime, storage and sharing capabilities differ
- If concerned about privacy, may need to be careful with any of them

Find full package list and comments at https://osf.io/9gfns/

Thank you to all package authors!

Plans for *simdata*

We plan to create a collection of data generators

From real-world datasets, and for useful artificial data

Follow active development at https://github.com/matherealize/simdata

What else is out there?

Packages for

- Survival data (specific models, often parametric)
- Counterfactual data (causal inference)
- Genetics (many packages)
- Structural equation models
- …