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

Michael Kammer, Medical University of Vienna

michael.kammer@meduniwien.ac.at



Synthetic data is <u>artificial data</u> that is generated from original data <u>and a</u> <u>model</u> 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

- * Measuring similarity can be done in many ways:
- by reproduction accuracy of a single intended application
- by domain angostic scores for dataset

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

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

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

-3.181471	0.1914619
-3.645042	0.1984420
-2.701850	0.1816263
-3.273654	0.1927918
-3.802442	0.1886332
	-3.645042 -2.701850 -3.273654



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 Heinze1Anne-Laure Boulesteix2Michael Kammer1,3Tim P. Morris4Ian R. White4on 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 Heinze1Anne-Laure Boulesteix2Michael Kammer1,3Tim P. Morris4Ian R. White4on behalf of the Simulation Panel of the STRATOS initiative

TUTORIAL IN BIOSTATISTICS

WILEY Statistics

Using simulation studies to evaluate statistical methods

Tim P. Morris¹ | Ian R. White¹ | Michael J. Crowther²

OPEN d ACCESS Freely available online

A Plea for Neutral Comparison Studies in Computational Sciences

Anne-Laure Boulesteix¹*⁹, Sabine Lauer¹⁹, 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



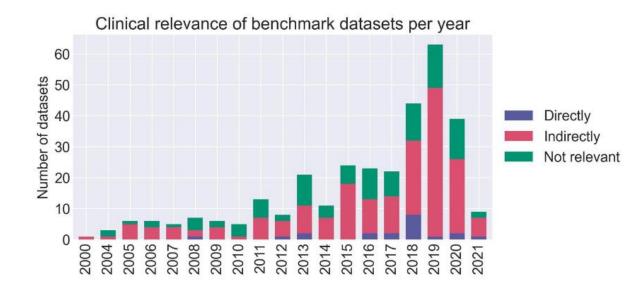
How to get data for simulation studies?

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

📕 .hg I .Rproj.user	01.10.2021 15:05 13.09.2022 16:49	Dateiordner Dateiordner	Data	a generator 1	<pre>CI</pre>	22
🧑 coxphfDev	01.10.2021 15:05	Dateiordner		<u>.</u>	Formation any feature intervention of the control of the cont	444 # dette umanie vector of coefficience 445 # dette et 445 # dette et 445 # dette dette dette dette dette dette dette
🧿 Data	01.10.2021 15:05	Dateiordner			12 . Super of roots	 Provide the structure is seen to seen to be an interest and the second se
Helper	01.10.2021 15:05	Dateiordner			1 Provide the second	62 Plength ML, Calability Chains the standard deviation or similar. 63 Plength ML, Intercapt 65 Plength Tocation of Intercapt In size, calability Mirst position, can be Mult, 65 Plan To Transage Is assumed.
5 scenarios	01.10.2021 15:05	Dateiordner				<pre>65 F Oreare 65 F / Oreare 65 F / Orearis 65 F / Orearis</pre>
-	01.10.2021 15:05	Dateiordner			and a second and a second and a second	462 d' variables bofore modeling, morare concer and scales from a scaled 668 d' variables active, "scaledorer" i Maragener, "scaledorer", "scaledorer", 668 unscale.cod = "mortionist, constra = mac, scales = macc, 669.
🥖 testing	16.11.2015 17:23	HGIGNORE-Datei	1 KB		<pre>/* (LawTOD) {</pre>	<pre>Figure 1 Figure 2 Figure 2 F</pre>
 .hgignore .Rhistory 	12.07.2023 16:16	RHISTORY-Datei	18 KB		Lower water services of the services and the services Lower services and the servi	01
					A destination instance provement (n) installed seals from the start of the start of the installed regulation in the start of the regulation - starting control (and regulation) (n) is a start of the start of the start of the start of the start of the start of the start of the	
📝 ex_Geb.R	30.07.2015 14:00	R-Datei	13 KB		<pre>iii reaction big ff</pre>	<pre>11 State of the state of t</pre>
📝 ex_OAK_data.R	29.06.2015 15:54	R-Datei	2 KB		Juniz Communication (Juniz) Additional (Juniz) Addition	<pre>book = book = book</pre>
💋 ex_prostate.R	23.07.2015 14:39	R-Datei	6 KB		<pre>/ ** first/subject/string/subje</pre>	802 41 602 1 603 1 604 1,000,2000,200 1000,000,000 1000,000,000 1000,000,000 1000,000,000 1000,000,000 1000,000,000,000 1000,000,000,000 1000,000,000 1000,000,000 1000
💇 ex_prostate_scaling.R	03.05.2018 17:20	R-Datei	7 KB			600) 602 m ² Ottile computation of highest density <u>invterval</u>
🛃 ex_vte.R	04.08.2016 10:07	R-Datei	10 KB		control and instantial architecture (and the control); control and instantial architecture(); control	* desarts to a seed of drars from bothog sayestaw desa weatysts by <u>writeches</u> . The seed of drars from bothog sayestaw desarts from a sample of representative values. The compares indukts descent from resval.
ex_vte_1000sim.rds	08.03.2016 15:44	RDS-Datei	651 993 KB		Predgatarda, Entremos, -) de de conservations -) de de conservations - (anternational de conservations) de conservations - (anternations) de conservatio	706 # Sampleved 707 # is a vector of representative values from a probability distribution.
🔬 ex_vte_cancer.R	22.07.2015 17:07	R-Datei	3 KB		<pre>#INTEL - MENDOLEMAN, - ME</pre>	To a comparison of the second
📝 ex_vte_cancer_data.R	22.07.2015 14:57	R-Datei	6 KB		101 - 1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 <t< td=""><td><pre>Terminal that is to activate. Terminal activation of the second se</pre></td></t<>	<pre>Terminal that is to activate. Terminal activation of the second se</pre>
📝 ex_vte_data.R	08.03.2016 15:18	R-Datei	3 KB		200 (1 201 (100,4000) 201 anotes(110,10, x, y + y, y,ul = y_un1), 201 (1) (1) (1) (1) (1) (1) (1) (1) (1) (1	773 -) 120 - Sorreaff - Sorreaffs[onto.min(cinitoto)] 120 - Sorreaffs[onto.min(cinitoto)] 121 - Sorreaffs[onto.min(cinitoto)] - cisikane] 122 - Sorreaffs - Sorreaffs, Sorreaffs
📝 helper.R	08.11.2015 18:33	R-Datei	3 KB		201 r un models 202 r un models 203 r un models 203 variation 203 variation 203 variation 204 r un models	992 - Den in 2012 - Concil and the concentration of the concentration of the concentration 2014 - Andrea and Concentration of the concentration of the concentration 2014 - Concentration and Concentration
📝 poster_computations.R	29.07.2015 22:33	R-Datei	1 KB		The second secon	725 F Coder Field to Be On Actual acts Ecse() 725 F 2005 Fields and coastate 735 F Topo: code add x_g3 argument to functions 737 model gm or _Mutching, y, family g gaussian",
📝 poster_graphics.R	06.03.2016 19:55	R-Datei	15 KB		11. If Character principal 22. A mainter with the section confidence intervals 23. A mainter predictions for training data	<pre>// Province and province or frequencies // Province and province or frequencies // Province and province or frequencies // Province and province and province or province</pre>
💼 r_template.docx	21.07.2015 15:56	Microsoft Word-D	16 KB		<pre></pre>	<pre>25 = = gin(y data = data.fram(y, y), fanty = fanty) 22 = f(fi.nut)(variable_meas) 23 = variable_meas) 24 = variable_meas = c(financeger", cohemes(x))</pre>
🕺 README.txt	09.02.2016 15:46	TXT-Datei	1 KB		041 041 <td>743 res = 1150) 744 # confriences 745 # confriences 746 # confriences</td>	743 res = 1150) 744 # confriences 745 # confriences 746 # confriences
📝 README_Rechner4.txt	09.02.2016 16:33	TXT-Datei	1 KB		2000 - 1000 (1000) (1000) 	<pre>control of the sector of</pre>
💕 report_template.docx	03.02.2016 16:32	Microsoft Word-D	18 KB			<pre>32 suppresentations(fet_confine(n_text) = c_l(ext))) 33 manas(res(cosf) = c(nos(), 'creat(n_text) = c_l(ext))) 33 manas(res(cosf) = c(nos(), 'creat(n_text) = los(n_t) = lo</pre>
shRinkage.Rproj	12.07.2023 16:15	R Project	1 KB		<pre>************************************</pre>	<pre>s protections relative = stra.Freedoweg1 = see. strategrad = stra.Freedoweg1 = see. strategrad = strategrad = strategrad = the '.' strategrad = strategrad = strategrad = strategrad = (strategrad = strategrad = strategrad = strategrad = strategrad = strategrad = strategrad = (strategrad = strategrad = strat</pre>
Simulation.docx	16.02.2016 17:24	Microsoft Word-D	1 135 KB	Data wawarata	Constructions and an analysis and analysis analysis analysis and analysis and analysis analysis and analysis and analysis and analysis analysis analysis and analysis analys	ini 1931 - 1 1941 - 1 1951 - Nosel_practe ← functionix, y, family = "severae".
simulation_graphics.R	07.03.2016 11:15	R-Datei	13 KB	Data generato	Totol Control Contr	<pre>most_practs <- mentors</pre>
simulation_helper.R	09.03.2016 14:39	R-Datei	93 KB	U	Active Section 2010 Storemouser, searce,	P22 m = stm(γ → , same - data.reas(bd, b i= 0, sing = resuld), γ), P22 P21
simulation_report.R	21.03.2017 17:15	R-Datei	30 KB		Contraction of the contract	777 ras = 11sc)
🖉 simulation_run.R	23.02.2016 11:13	R-Datei	4 KB		202 # comprises to is comparable lacks without being inflation 203 # by dest reconnected 204 run((b)) comprises result of the reconnected reconnecte	restore - set. Prestoret - set. restore - set. Prestoret - set. restoret - set. rest. restoret - set. restoret - set. restoret - set.
🖉 talk_Goettingen.R	11.03.2016 15:31	R-Datei	11 KB		30 run(1) [loger] biosci [sci] sci] 300 - 36 6.634_Seci [run(10] [loger] transmission 370 run(10] [loger] transmission 381 run(10] [loger] transmission 382 run(10) [loger] transmission	<pre>3 . 17 (Suppose - portial) (</pre>
tmp.docx	30.07.2015 20:48	Microsoft Word-D	17 KB		act asl_cat(init(i))isoffacer rem(i)isoffacer(init(i))isoffacer rem(i)isoffacer(init(i))isoffacer(i)) remote the remote the remote the remote the remote the remote remote the remote the remote the remote the remote the remote remote the remote the	resigned - data.frame(node) - name. 192 resigned - data.frame(node) - name. 293 est_Theat_predictor - predict(m, newspa - x.
imp.docx	14.07.2015 14:25	R-Datei	1 KB		<pre>zrr # change inclusion of 28 # write in inclusion prove and dowrange 28 # offer inclusion prove and as if we constant other on b 28 # offer inclusion inclusion comparable 28 # offer inclusion inclusion comparable</pre>	756 (1990) 757 (1990) 758 (1990) 758 (1990) 759 (1990)
			1100		and the second s	<pre>rest_listing.fit = fuertexit, p. fuert_rist_rist_rist_rist_rist_rist_rist_ri</pre>





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

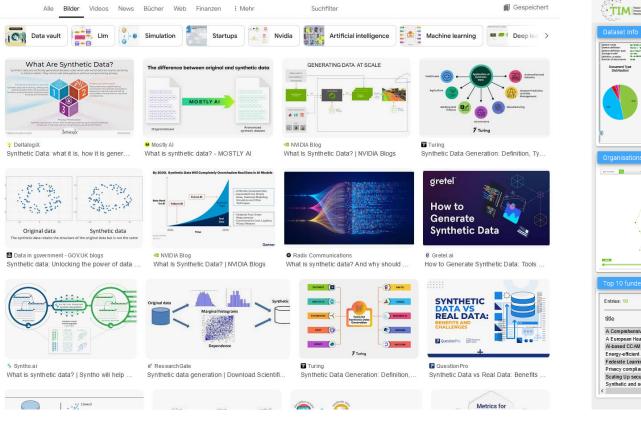
But...

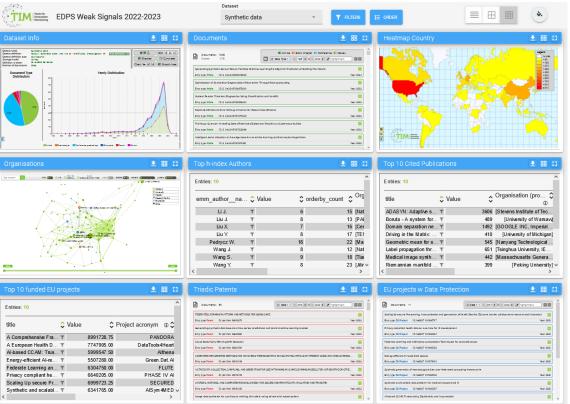
- Avoid oversimplification and monopolization
- Dissemination
- Creating and documenting reliable code takes a lot of time!



A first step: what is out there?

Synthetic data is a hot topic





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

Michael Kammer



Google search "Syntethic data" on 8.1.2025

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
- ← → C
 A https://cran.r-project.org/web/packages/available_packages_by_name.html
 Available CRAN Packages By Name
 <u>ABCDEFGHIJKLMNOPQRSTUVWXYZ</u>
 A3
 Accurate, Adaptable, and Accessible Error Metrics for Predictive Models
 Conditional Aalen-Johansen Estimation
 <u>AATtools</u>
 Reliability and Scoring Routines for the Approach-Avoidance Task
 <u>ABACUS</u>
 abasequence
 Coding 'ABA' Patterns for Sequence Data

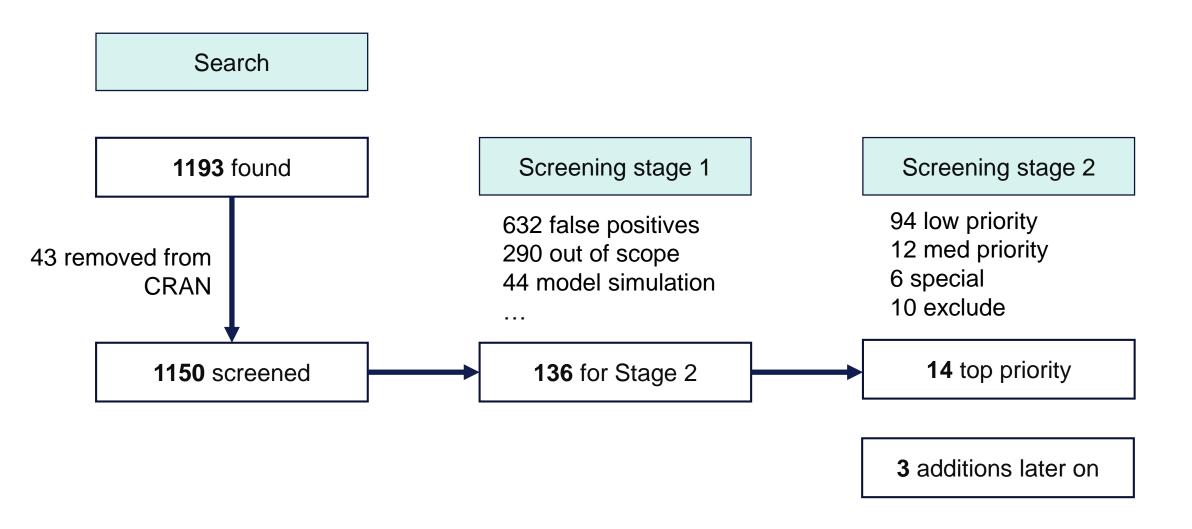
Keywords:

*simulat** (includes terms such as simulation, simulate, simulating...) *generat**

synth*

Search date: 18.07.2023 with subsequent updates

Search and screening results





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

	Туре	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



Power polynomials

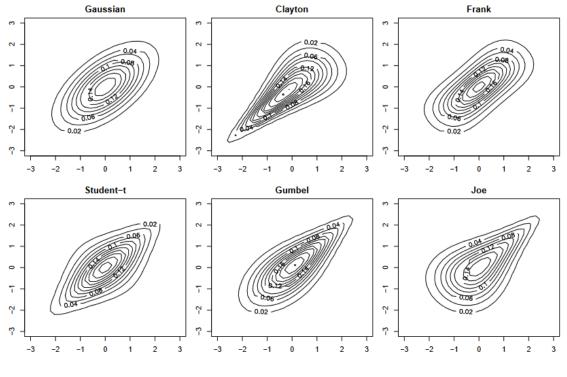
- Specify non-normal continuous data as $X = a + bZ + cZ^2 + dZ^3$, $Z \sim N(0,1)$ (Fleishman 1978)
- Extended to multivariate data (Vale 1983)
- Extended to further moments (Headrick 2002)
- Long standing in social science and psychometrics (Headrick 2010)
- Implementations in PoisBinOrdNonNor, SimCorrMix,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



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).



Normal to anything (NORTA, Cario 1997)

• Given distribution functions $F_i(s) = P(X_i \le s)$ and correlation matrix Σ_X :

Generate multivariate standard normal vectors Z with correlation Σ_Z

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

• To achieve desired final correlation matrix, Σ_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.



Model based (conditional specification)

- Approximate joint distribution by sequence of conditional distributions (models) Start by synthetizing x₁' from x₁ (i.e. sample with replacement) Fit model for x₂ using observed x₁ and sample from fitted model to obtain x₂' 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).



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



	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

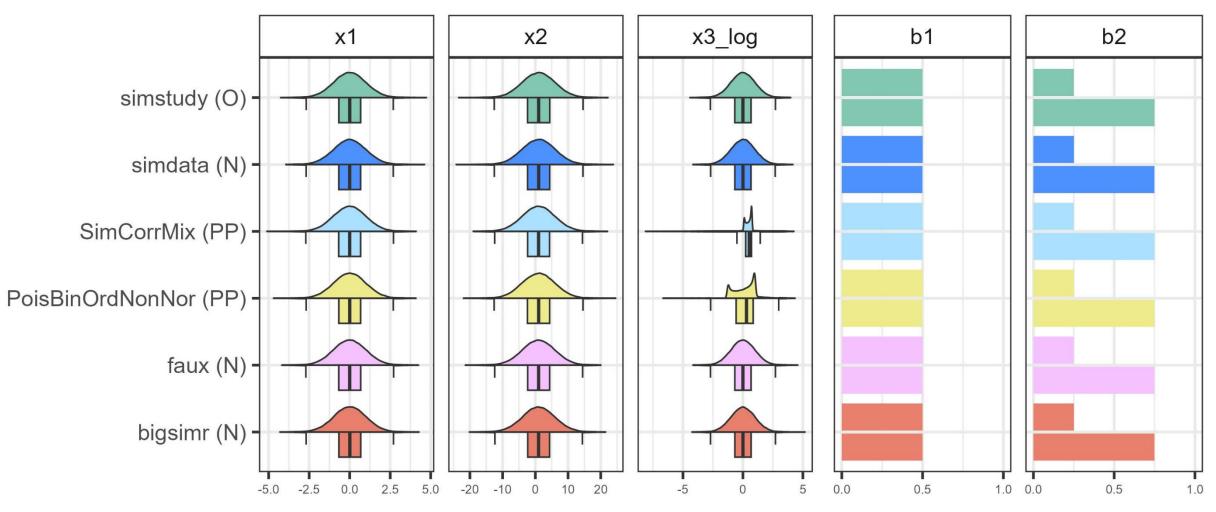
- 5 variables: N(0, 1), N(1, 5), LogN(0, 1), Bern(0.5), Bern(0.25),
- Target pairwise Pearson correlation: 0.2
- 100,000 observations

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)

	bigs	imr	(N)				faux	: (N))				Pois	Bin	Ord	Nor	Nor	(PP)
b2	0	0	0	0	0	b2	-0.01	0.04	-0.01	-0.05	0	b2	0	-0.01	-0.19	0	0	
b1	0	0	-0.01	0	0	b1	0	0.02	0	0	-0.05	b1	0	0	-0.19	0	0	
х3	-0.01	-0.01	0	-0.01	0	x3	0.01	0	0	0	-0.01	х3	-0.02	-0.02	0	-0.19	-0.19	
x2	0	0	-0.01	0	0	x2	0	0	0	0.02	0.04	x2	0.02	0	-0.02	0	-0.01	
x1	0	0	-0.01	0	0	x1	0	0	0.01	0	-0.01	x1	0	0.02	-0.02	0	0	
	4	2	\$	à	5		4	2	3	à	52		+	2	3	à	52	
	Sim				^{>})		sime					3	simo					
b2	Sim)	b2	sim	stud	y (C			b2	simo 0				0	
		Cor 0	rMix	(PF 0			sim: -0.05	stud -0.05	y (C)) -0.08		b2		data	(N)		-	
b2	0	Cor 0	rMi× 0.01 -0.02	(PF 0	0	b2	sim: -0.05 -0.04	stud -0.05	y (C -0.08 -0.08)) -0.08 0	0	b2 b1	0	data ₀	(N) 0	0	0	
b2 b1	0	Cor o	rMi× 0.01 -0.02	c (PF 0 0	0	b2 b1	sim: -0.05 -0.04	-0.05 -0.04	y (C -0.08 -0.08 0)) -0.08 0 -0.08	0 -0.08	b2 b1 x3	0	data o o	(N) 0	0	0	
b2 b1 x3	0 0 0	o o o	0.01 -0.02 0	0 0 -0.02	0 0 0.01	b2 b1 x3	sims -0.05 -0.04 -0.04	-0.05 -0.04 -0.05	y (C -0.08 -0.08 0 -0.05	-0.08 0 -0.08 -0.04	0 -0.08 -0.08	b2 b1 x3 x2	0 0 0	o 0 0	0 0 0	0 0 0	0 0 0	



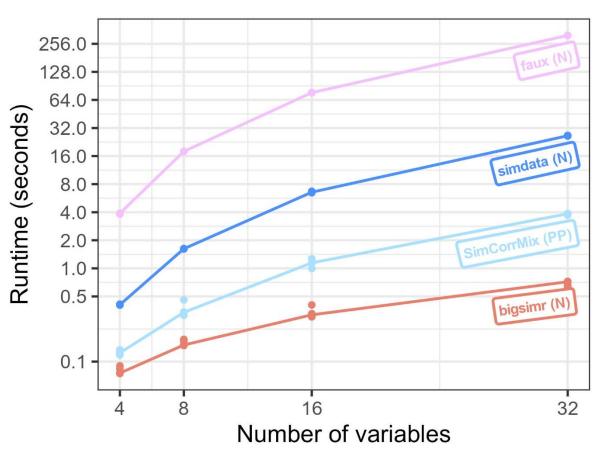
Artificial data comparison: runtime scaling

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





28

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

							<u> </u>
BPdia	-0.13	0	0.07	0.16	0.09	0.35	1
BPsys	-0.06	0.48	0.04	0.1	0.12	1	0.35
BMI -	0.05	0.01	-0.11	0.89	1	0.12	0.09
Weight	-0.25	-0.05	-0.06	1	0.89	0.1	0.16
Race	0	0.01	1	-0.06	-0.11	0.04	0.07
Age	-0.02	1	0.01	-0.05	0.01	0.48	0
Gender	1	-0.02	0	-0.25	0.05	-0.06	-0.13
G	pender	P.96	Race 1	Neight	BMI	apeys	BPdia

Michael Kammer

Simulation packages in R

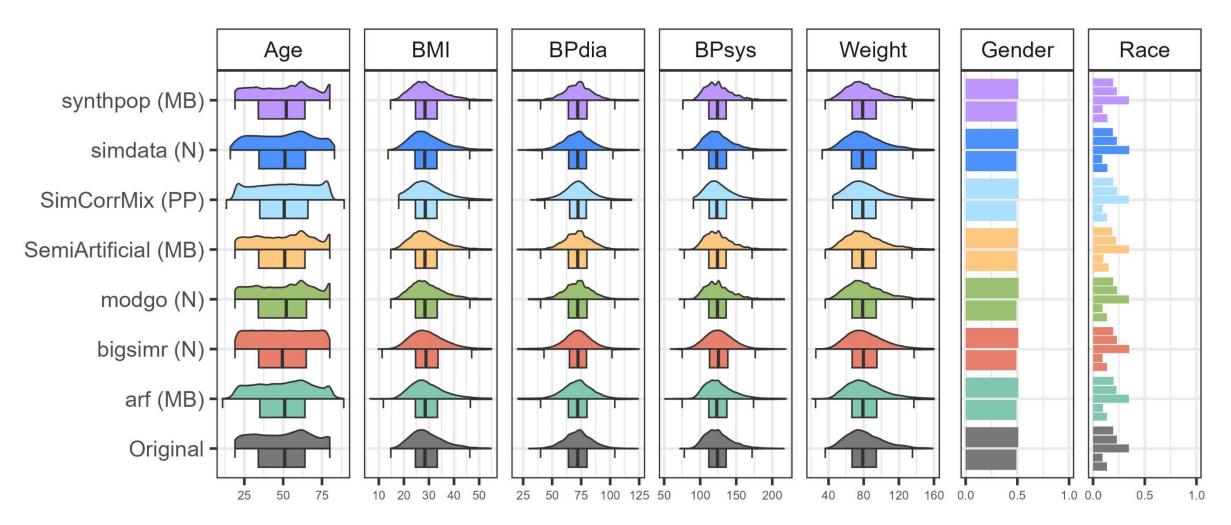


Specification of data generators quite different

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)

BPdia	0.01	0	0	0	0	0	0
BPsys	0	0	0	0	0	0	0
BMI	0	0	0	0	0	0	0
Weight	0	0	0	0	0	0	0
Race	0	0	0	0	0	0	0
Age	0	0	0	0	0	0	0
Gender	0	0	0	0	0	0	0.01

modgo (N)

BPdia	0	-0.04	0.01	0	0.01	0.04	0
BPsys	-0.04	0.03	0	0.03	0.03	0	0.04
BMI	-0.03	0	-0.02	-0.02	0	0.03	0.01
Weight	-0.02	-0.02	-0.02	0	-0.02	0.03	0
Race	0	0	0	-0.02	-0.02	0	0.01
Age	0.01	0	0	-0.02	0	0.03	-0.04
Gender	0	0.01	0	-0.02	-0.03	-0.04	0

SimCorrMix (PP)

BPdia	0	0.08	-0.08	0	0	0	0
BPsys	0	-0.45	0.44	0	0	0	0
BMI	0	-0.12	0.12	0	0	0	0
Weight	0	-0.01	0.01	0	0	0	0
Race	-0.02	0	0	0.01	0.12	0.44	-0.08
Age	0.02	0	0	-0.01	-0.12	-0.45	80.0
Gender	0	0.02	-0.02	0	0	0	0
	5	0	0.	×	10	5	.0

synthpop (MB)

01	-0.01 0		-0.01 -0.01	0.01	-0.01	
	0	0	-0.01	0	0	
			0.01	0	0	-0.01
)	0	0	0	0	0	0.01
)	-0.01	0	0	0	-0.01	-0.01
01	0	0	0	0	0	-0.01
)	0	0	-0.01	0	0	-0.01
)	0	-0.01	0	0	0.01	0.03
) 01)) -0.01 01 0) 0	-0.01 0 01 0 0 0 0 0	-0.01 0 0 01 0 0 0 0 0 0 0	-0.01 0 0 0 01 0 0 0 0 0 0 0 0 0	-0.01 0 0 0 -0.01 01 0 0 0 0 0 01 0 0 0 0 0 0 0 0 -0.01 0 0



simdata (N)

BPdia	0.01	0	0	0	-0.01	0.01	0
BPsys	0	-0.01	0	0	0	0	0.01
BMI	0	0	-0.01	0	0	0	-0.01
Weight	0	0	-0.01	0	0	0	0
Race	0	0	0	-0.01	-0.01	0	0
Age	0.01	0	0	0	0	-0.01	0
Gender	0	0.01	0	0	0	0	0.01

SemiArtificial (MB)

BPdia	-0.01	0	-0.01	0	0	0	0
BPsys	0	0	-0.01	0	-0.01	0	0
BMI	-0.01	-0.01	0	-0.03	0	-0.01	0
Weight	0	0	0	0	-0.03	0	0
Race	0	0	0	0	0	-0.01	-0.01
Age	-0.02	0	0	0	-0.01	0	0
Gender	0	-0.02	0	0	-0.01	0	-0.01

arf (MB)

BPdia	0	0	-0.01	0	0	-0.05	0
BPsys	0	-0.03	0	0	-0.01	0	-0.05
BMI	-0.03	0	0.01	-0.09	0	-0.01	0
Weight	0.04	0.01	-0.01	0	-0.09	0	0
Race	0	0	0	-0.01	0.01	0	-0.01
Age	0	0	0	0.01	0	-0.03	0
Gender	0	0	0	0.04	-0.03	0	0



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

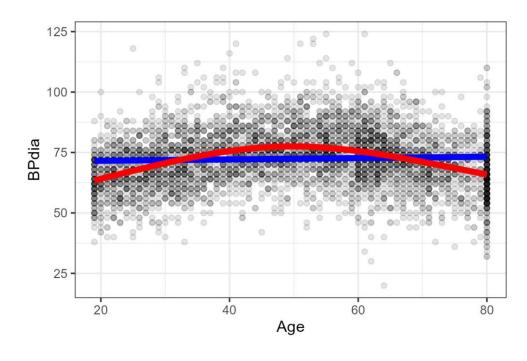
All results in seconds, medians of 10 runs, Windows 10, Intel i7-10700K@3.8Ghz

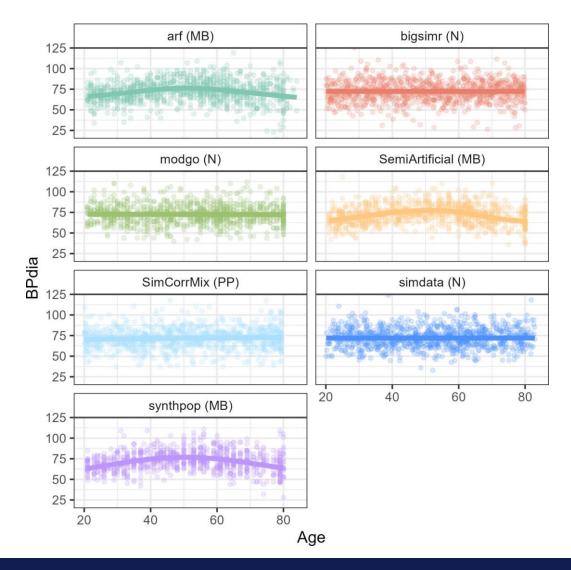


Real world data: non-linearities

Correlations imply form

Methods based on correlations restrict form of relations







Real world data: distance correlations

bigsimr (N)

BPdia -0.02 -0.15 0.02 0.01 0.01 -0.08 0

BPsvs -0.07 -0.02 -0.01 -0.03 -0.01 0 -0.08

Weight -0.02 0 -0.02 0 0.01 -0.03 0.01

Age Gender

BMI -0.01 -0.05 0.01 0.01 0 -0.01 0.01

Race 0.01 -0.08 0 -0.02 0.01 -0.01 0.02

-0.02 0 -0.08 0 -0.05 -0.02 -0.15

0 -0.02 0.01 -0.02 -0.01 -0.07 -0.02

det de ce drit M y dia

Difference

achieved - observed

Distance correlation measures also non-linear relations.

	synthpop (MB)						simdata (N)								Se	m	iAi	tifi	cia	al	(MB)	arf (MB)										
BPdia	-0.02	0.01	-0.01	0	0.03	0.01	0	BPdia	-0.03	3 -0.	16 -0.	03 0.	.01 -0	.01 -0	0.07	0	BPdia	0.01	-0.02	-0.01	0.04	0.04	-0.0	2 0	BPdia	-0.01	-0.02	0	0	0.02	-0.02	. 0
BPsys	-0.02	0	0	-0.01	0	0	0.01	BPsys	-0.04	4 -0.	03 -0.	02 -0	.04 -0	.03	0 -0	0.07	BPsys	-0.01	0.02	0.01	0	-0.0	2 0	-0.02	BPsys	0	-0.04	0.02	0	-0.02	0	-0.02
BMI	-0.01	0.01	0	0.01	0	0	0.03	BMI	0.02	2 -0.	05 -0.	01 0.	.01	0 -0	0.03-0	0.01	BMI	-0.01	0.03	0.02	-0.02	2 0	-0.0	2 0.04	BMI	0.02	0.02	-0.01	-0.08	0	-0.02	0.02
Weight	0.02	0.03	-0.01	0	0.01	-0.01	0	Weight	-0.04	4 -0.	03 -0.	06	0 0.	01 -0	0.04 0	0.01	Weight	0.03	0.01	0.02	0	-0.0	2 0	0.04	Weight	-0.03	0.03	0	0	-0.08	0	0
Race	0.02	-0.01	0	-0.01	0	0	-0.01	Race	0.03	8 -0.	07 C) -0	.06 -0	.01 -0).02-0	0.03	Race	0.02	-0.02	0	0.02	0.02	0.0	1 -0.01	Race	0.01	-0.02	0	0	-0.01	0.02	0
Age	-0.01	0	-0.01	0.03	0.01	0	-0.01	Age	0	C	-0.	07 -0	.03 -0	.05 -0	0.03-0	0.16	Age	-0.01	0	-0.02	0.01	0.03	0.0	2 -0.02	Age	0	0	-0.02	0.03	0.02	-0.04	-0.02
Gender	0	-0.01	0.02	0.02	-0.01	-0.02	2-0.02	Gender	0	C	0.0	03 -0	.04 0.	02 -0	0.04 -0	0.03	Gender	0	-0.01	0.02	0.03	-0.0	1-0.0	1 0.01	Gender	0	0	0.01	-0.03	0.02	0	-0.01

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

modgo (N)

BPdia 0.01 -0.16 0 0.01 0.01 -0.05 0

BPsvs -0.02 0.04 0 0.03 0.05 0 -0.05

Weight 0.03 -0.05 -0.05 0 0.01 0.03 0.01

Race

Gender

BMI -0.05 -0.02 -0.03 0.01 0 0.05 0.01

Age 0.01 0 -0.09 -0.05 -0.02 0.04 -0.16

0 -0.09 0 -0.05 -0.03 0 0

0 0.01 0 0.03 -0.05 -0.02 0.01

det de ce dit Mi ye dia



Michael Kammer Simulation packages in R

SimCorrMix (PP)

0 0.03 -0.09 0.01 0 -0.03 -0.03

-0.03 0 -0.03 0 0.01 -0.06 -0.04

0 -0.08 0 -0.03 -0.09 0.41 -0.05

-0.02 0 -0.08 0 0.03 -0.44 -0.12

det de ce drit mi sy dia

0 -0.02 0 -0.03 0 -0.08-0.01

 BPdia
 -0.01
 -0.12
 -0.05
 -0.04
 -0.03
 -0.1
 0

 BPsys
 -0.08
 -0.44
 0.41
 -0.06
 -0.03
 0
 -0.1

BMI

Weight

Gender

Race

Age

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

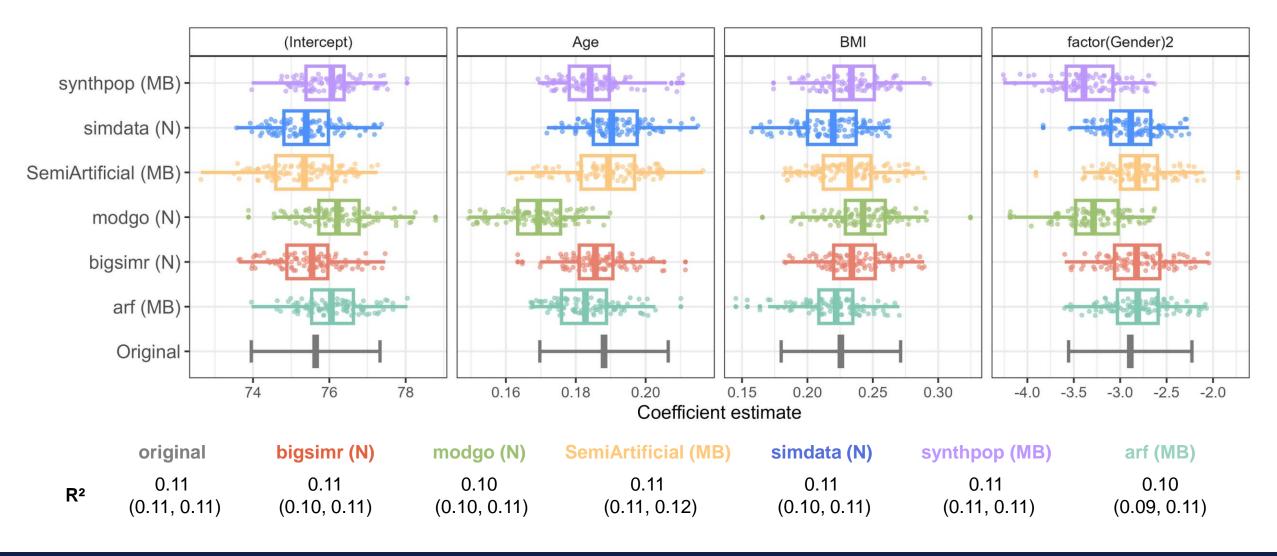
Outcome diastolic blood pressure

Covariate Age (3-knot restricted cubic spline)

Both models re-created 100 times.

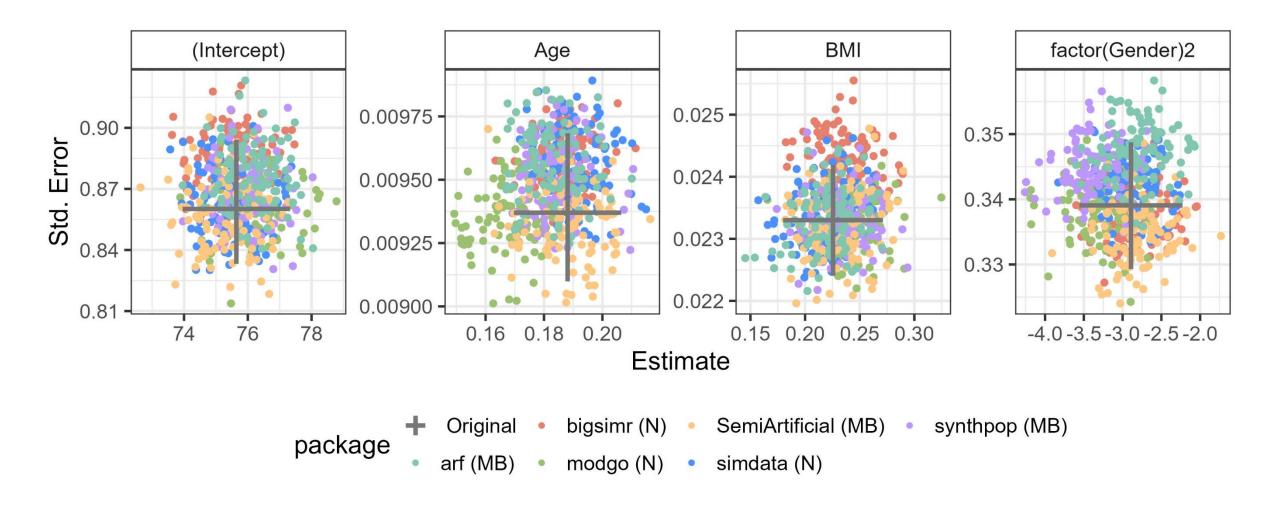


Recreating regression analyses: Model 1



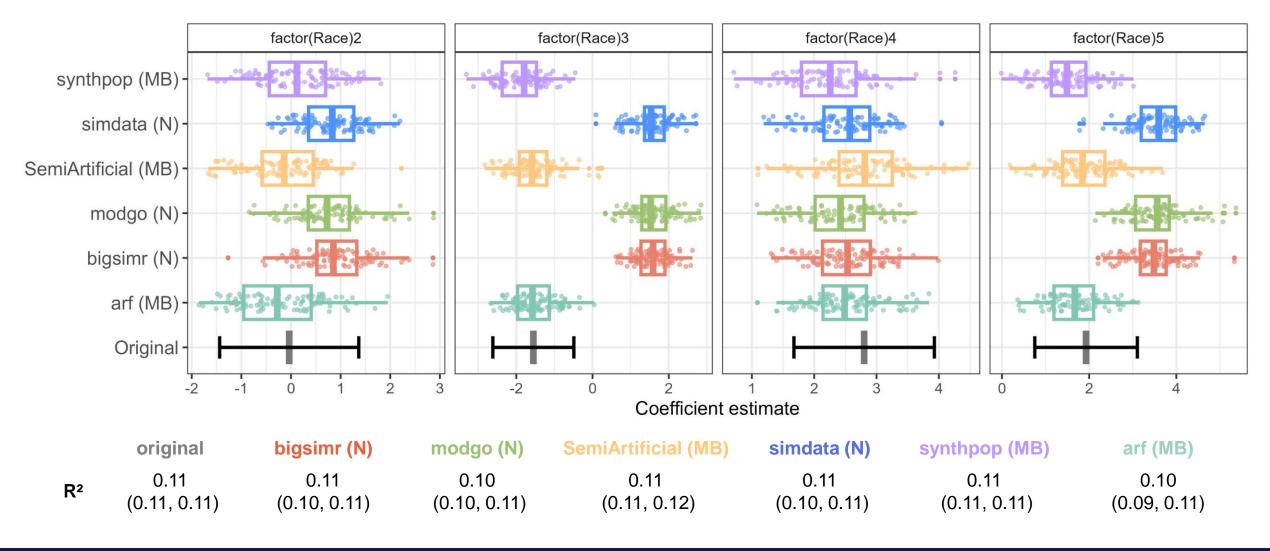


Recreating regression analyses: Model 1



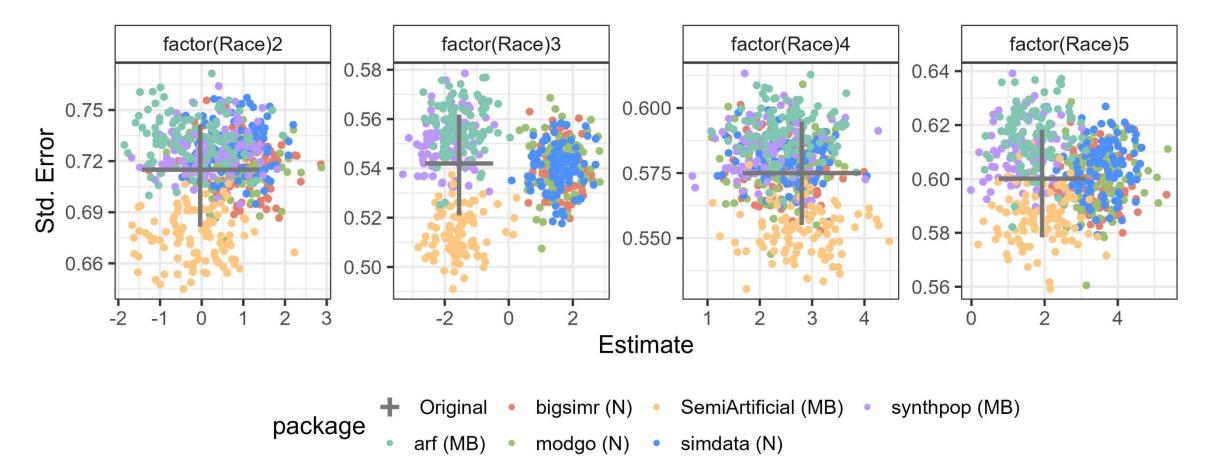


Recreating regression analyses: Model 1 addon





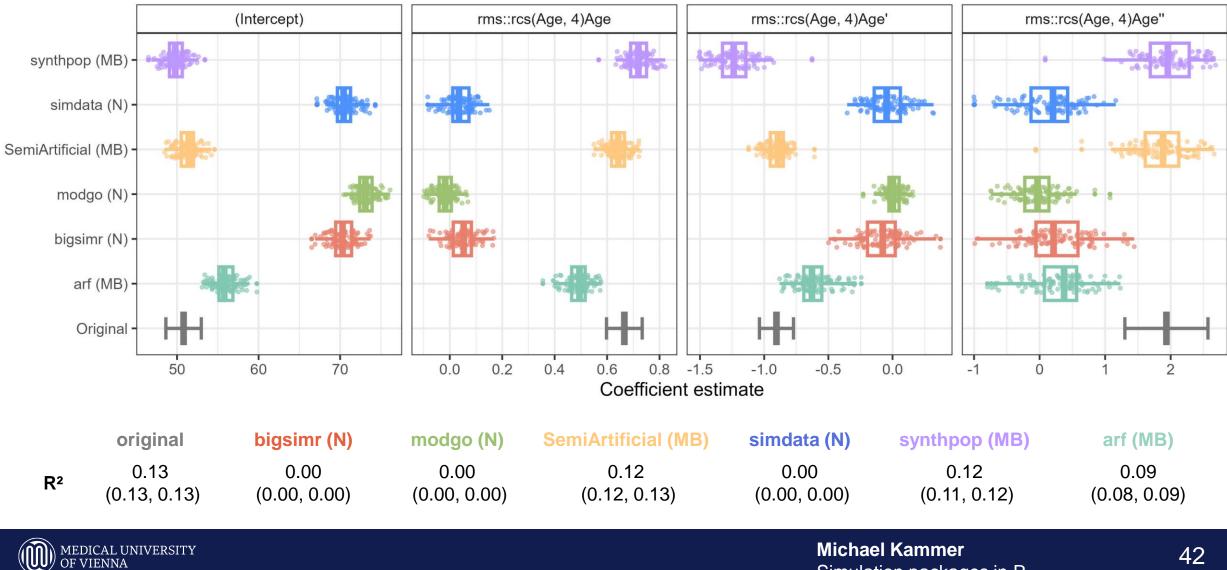
Recreating regression analyses: Model 1 addon



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

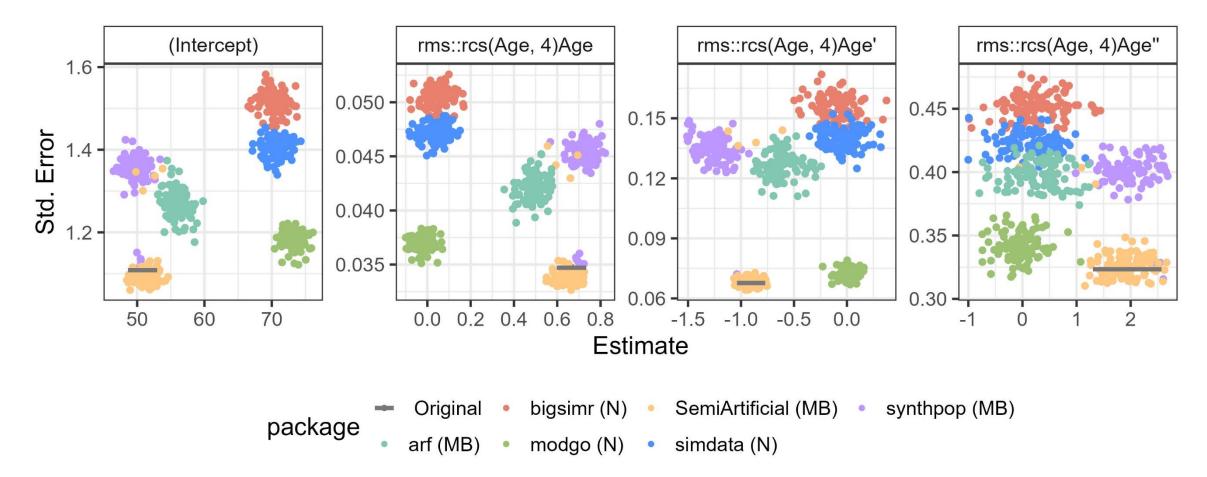


Recreating regression analyses: Model 2



Simulation packages in R

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	ОК	OK-Slow	Slow	OK	Slow
Privacy	OK	? (ECDF)	OK	? (RF)	? (ECDF)	?
MEDICAL UNIVERS	ITY			Michae	el Kammer	ЛЛ



Michael Kammer Simulation packages in R

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





Michael Kammer Simulation packages in R

What else is out there?

Packages for

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



. . .