A Data Simulation Tool for Evaluating Machine Learning Performance for Classification Problems

Matthew C. Myers, M.Ed.
University of Delaware

Abstract

This research presents a data simulation tool designed to provide large-scale meaningful data for classification algorithms (e.g., random forests). Using Python's `pandas`, `NumPy`, `SciPy`, and `random` libraries, each simulation generates an unlimited number of unique datasets derived from user-specified global parameters, which include constraints for sample size, the numbers of predictors and the degree of statistical noise, as well as the range of distributional characteristics for the predictors. Sample size is partitioned into two or more categorical dependent outcomes. Dataset characteristics are defined via `random`'s pseudo-random number generation. To emulate the natural covariation among predictors in real datasets, the tool transforms initial unadjusted predictor values using randomly specified covariance matrices. This research first details the algorithm. It then demonstrates its potential uses.

Introduction

- **Machine learning (ML)** performance can be evaluated with real and simulated data. Real data are often insufficient due to inbuilt limitations or excessive model demands (van der Ploeg et al., 2014).
- A **standard approach to data simulation** involves the researcher specifying parameters with tight control given their operational knowledge of the **target system** (Kim et al., 2017).
- This approach is sufficient for investigations centered on the target systems (e.g., medicine, geology, education; see Schulz et al., 2017). However, the limited breadth of varied datasets **precludes certain investigations** specifically related to evaluating and optimizing algorithm performance.
- We propose that **large-scale data assemblages** defined by a global parameter space encompassing multiple target systems are integral to comprehensive ML performance evaluation and optimization.

Purpose & Research Questions

Primary Purpose: To provide methodologists a tool for exploring ML research questions that demand large-scale data assemblages.

Secondary Purpose: To provide an educational tool for those learning about supervised ML algorithms.

Research Question 1: Can the algorithmic logic of this approach to simulation capably generate an assemblage of datasets defined by a global parameter space?

Research Question 2: To what extent do these simulated data emulate real data from real systems?

Logic

USER FRONT-END

0. Specify global parameter constraints.

EXECUTE PROGRAM

GENERATE PARAMETERS

1. For each dataset m_i randomly generate:
Total sample size n
Number of possible outcomes q
Number of meaningful predictors p
Degree of statistical noise z

PARTITION OUTCOMES

2. For each outcome q_j in dataset m_i:
Partition outcome q_1 to random proportion of sample size n_i
2a. If* q = *2:
Partition outcome q_2 to remaining sample $(n_i - q_1)$
2b. Else if* q *> 2:
Partition outcome q_2 to random proportion of remaining sample $(n_i - q_1)$
Partition outcome q_j to remaining sample $(n_i - q_{1+2+\ldots+[j-1]})$,
such that $q_1 + q_2 + \ldots + q_j = n_i$

SPECIFY UNCORRELATED PREDICTORS

3. For each predictor p_k in dataset m_i randomly select:
Distribution type (e.g., Normal, Bernoulli, Poisson)
Parameters (e.g., μ, σ, ρ, λ) for outcome q_1 (i.e., reference group)

4. Assign values to p_k for outcome q_1

5. For each additional outcome q_j:
Differentially adjust parameters based on reference group

6. Assign values to p_k for outcomes $q_2, q_3 \ldots q_j$

SPECIFY CORRELATION MATRIX

7. For dataset m_i determine size S of matrix M, where:

$$S_m = p_m \times p_m$$

8. Generate empty matrix M of size S:

$$M = \begin{pmatrix} p_1 & p_2 & p_3 & \cdots & p_k \\ p_2 & \emptyset & \emptyset & \emptyset & \emptyset \\ p_3 & \emptyset & \emptyset & \emptyset & \emptyset \\ \cdots & \emptyset & \emptyset & \emptyset & \emptyset \\ p_k & \emptyset & \emptyset & \emptyset & \emptyset \end{pmatrix}$$

9. The number of unique correlation coefficients r is given by:

$$r = \frac{p(p+1)}{p}$$

10. For each correlation coefficient r_i:
Generate a random value between -1.0 and 1.0

11. Assign each r_i to matrix M:

$$M = \begin{pmatrix} r_0 & r_1 & r_2 & \cdots & r_{i-3} \\ r_1 & r_0 & r_3 & & r_{i-2} \\ r_3 & r_2 & r_0 & & r_{i-1} \\ \cdots & & & & r_i \\ r_{i-3} & r_{i-2} & r_{i-1} & r_i & r_0 \end{pmatrix}$$

12. For matrix M, check all eigenvalues to verify positive-definiteness:
12a. If M is ***not*** positive-definite, then ***repeat steps 8-12***
12b. If M is positive-definite, then ***proceed to step 13***

TRANSFORM PREDICTORS

13. For each uncorrelated predictor p_k in dataset m_i:
Obtain correlated predictor x_k via Cholesky decomposition,
such that $x_1 = p_1$ and $x_k = rp_1 + \sqrt{1 - r^2}p_k$

GENERATE STATISTICAL NOISE

14. For each noise variable z_k in dataset m_i:
Assign values based on RNG indiscriminately to z_k for all outcomes q_j

LOOP UNTIL *m* DATASETS ARE GENERATED

Sample Simulation

Parameter specification by user for sample simulation

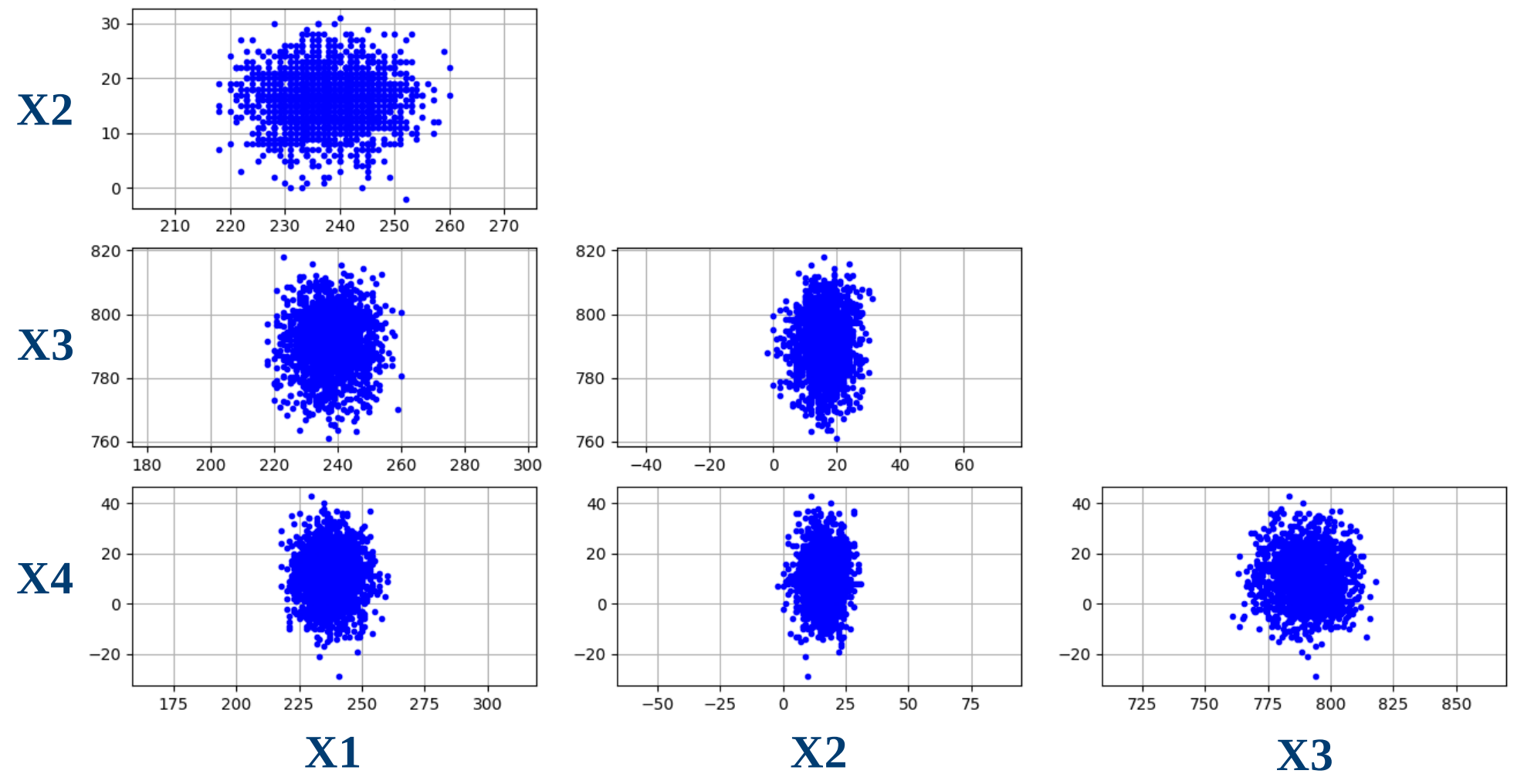
- total_datasets = 1e5
- max_decimals = 2
- min_n = 100
- max_n = 3000
- min_mu = 0
- max_mu = 800
- min_outcomes = 2
- max_outcomes = 3
- min_predictors = 0
- max_predictors = 10
- min_noise = 0
- max_noise = 10
- partition_ceiling = .75
- partition_floor = .25
- min_mean_diff = .999
- max_mean_diff = 1.001
- min_p = 0.0
- max_p = 1.0

Parameters for 10 randomly generated datasets

	Sample Size	Outcomes	Outcome Partition			Predictors	Noise
ID	n	q	n_{q1} (%)	n_{q2} (%)	n_{q3} (%)	p	z
1	1765	3	858 (.49)	419 (.24)	488 (.27)	4	1
2	2925	2	1184 (.40)	1741 (.60)	-	0	9
3	2607	2	1679 (.64)	928 (.36)	-	6	10
4	701	3	408 (.58)	158 (.23)	135 (.19)	10	10
5	935	2	281 (.30)	654 (.70)	-	4	10
…	…	…	…	…	…	…	…
9996	1256	2	498 (.40)	758 (.60)	-	7	2
9997	1863	3	641 (.34)	476 (.26)	746 (.40)	8	8
9998	2086	2	844 (.40)	1242 (.60)	-	5	2
9999	291	2	130 (.45)	161 (.55)	-	6	0
10000	842	3	608 (.72)	89 (.11)	145 (.17)	3	6

FOR SAMPLE DATASET 1

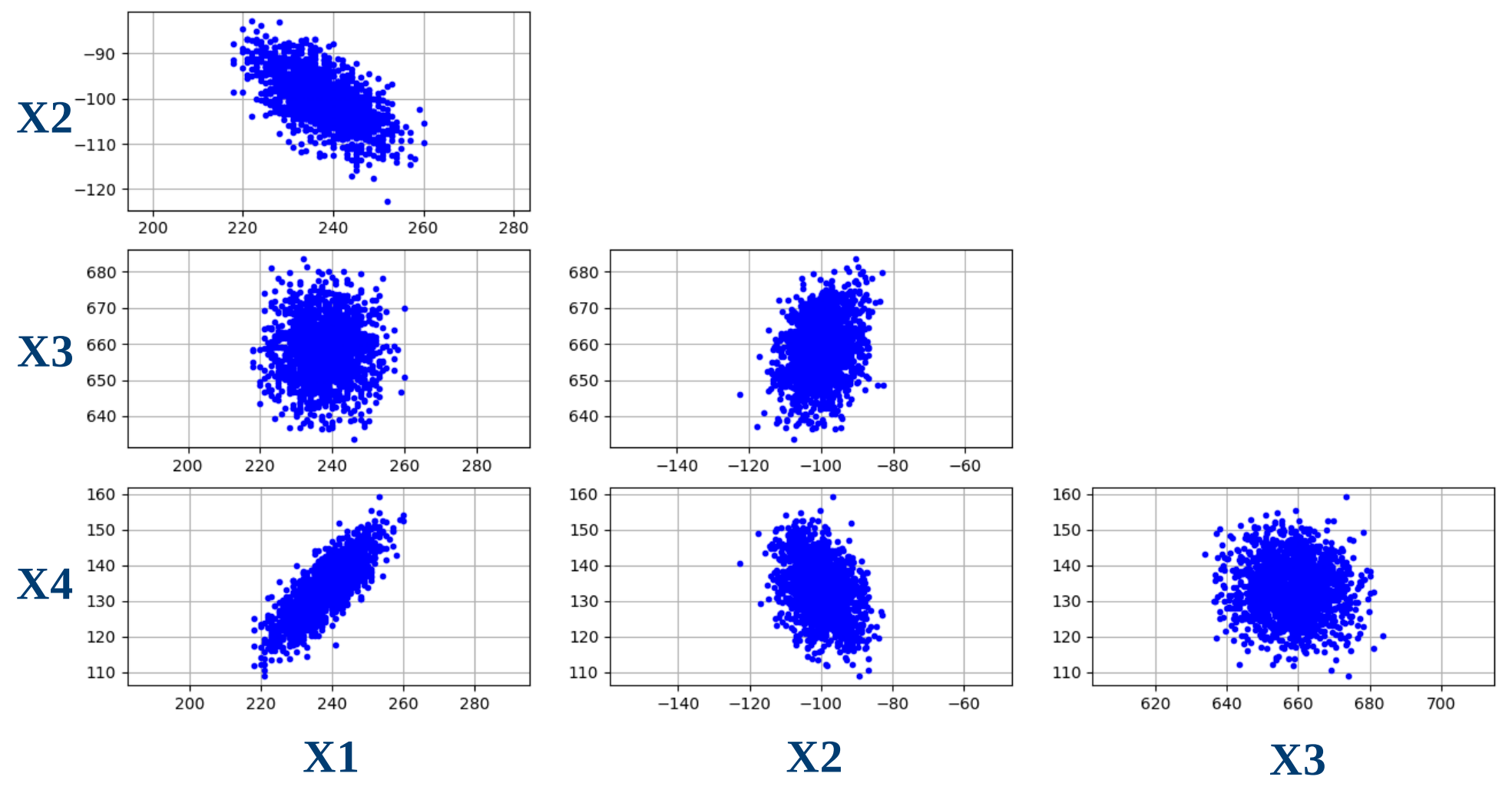
Randomly Generated Uncorrelated Predictors



Correlation Coefficients for Uncorrelated Predictors

Predictors	X1	X2	X3	X4
X1	1.000			
X2	.006	1.000		
X3	.019	.035	1.000	
X4	-.033	.030	-.013	1.000

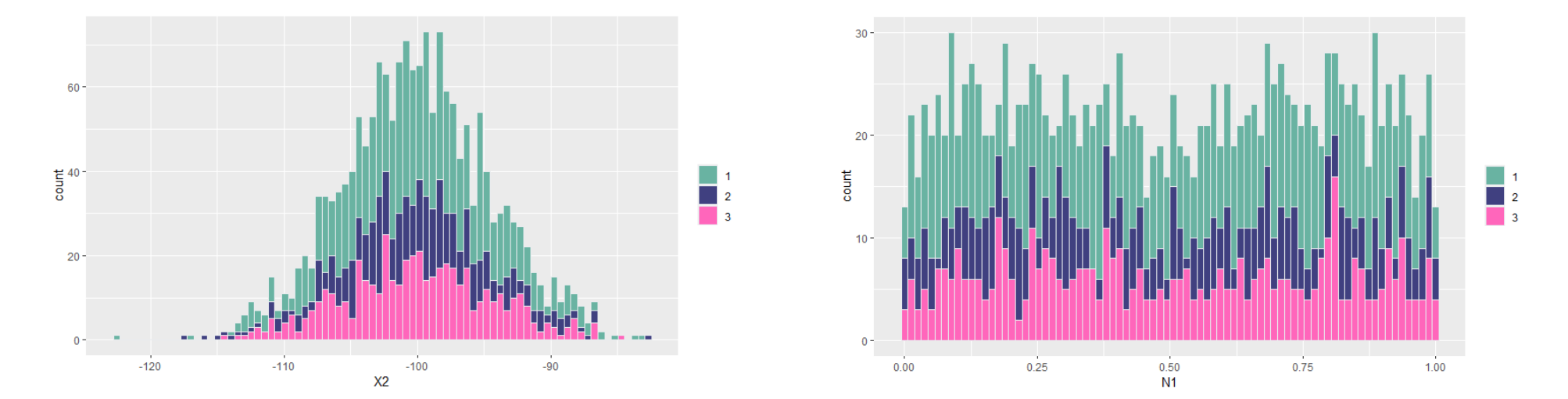
Correlated Predictors After Cholesky Decomposition



Correlation Coefficients for Transformed Predictors

Predictors	X1	X2	X3	X4
X1	1.000			
X2	-.608	1.000		
X3	.019	.292	1.000	
X4	.805	-.368	-.049	1.000

Comparative Distributions: Predictor vs. Noise by Group



Usage Example

The following example of the usefulness of this data simulation tool is a cross-section of a study currently in progress:

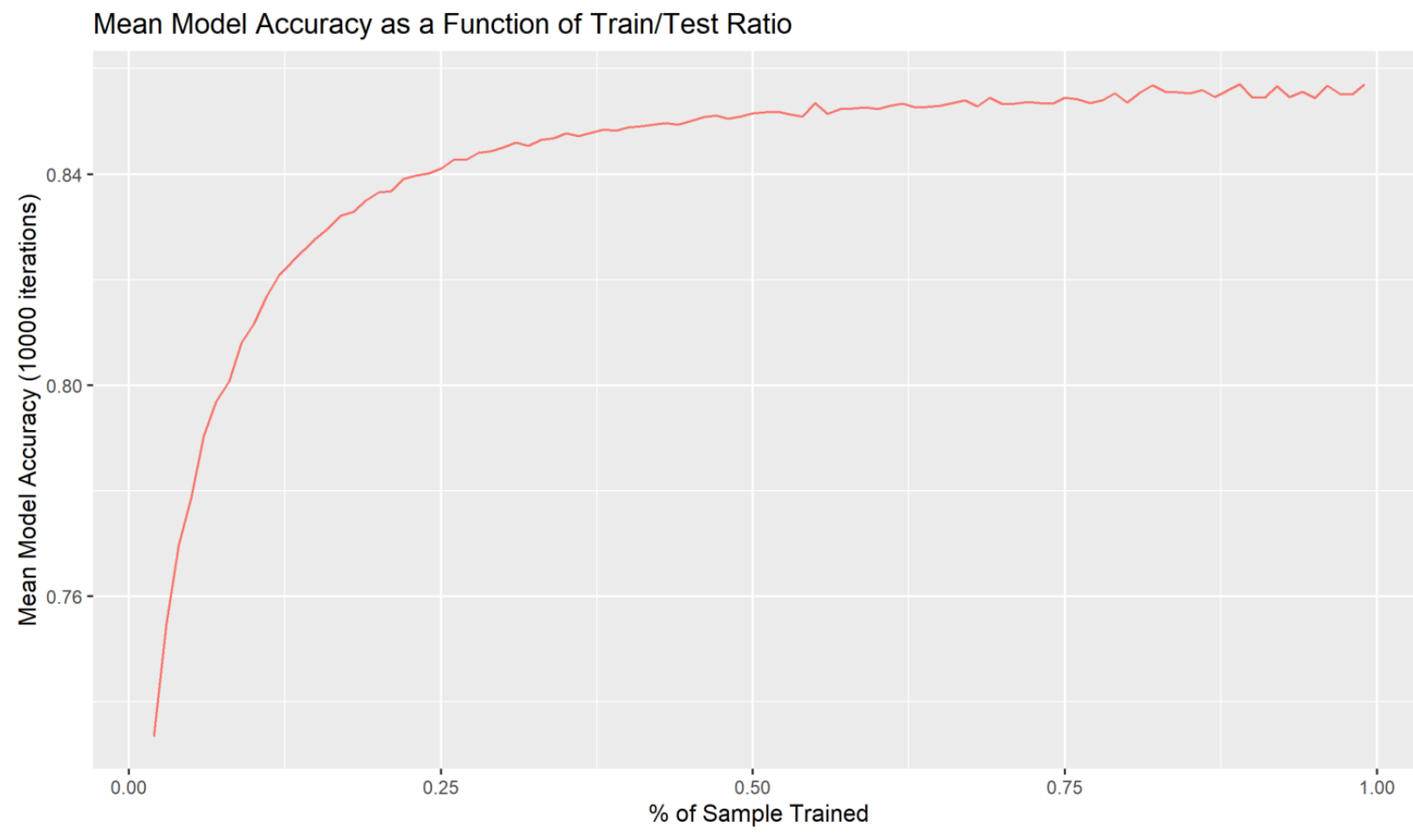
Research Questions

1. To what extent can the predictive performance of random forest classifiers be improved by optimizing the train/test (t/t) ratio specification procedure for a given dataset?
2. Is it possible to validate the Pareto Principle (i.e., 80/20 rule) that is widely used in classification problems?
3. Does there exist an average optimal t/t ratio parameter for the infinite set of all hypothetical datasets?

Rationale for large-scale data simulation

Research questions 2 & 3 cannot be investigated with standard, small-scale approaches to data simulation. They require an immense vector of unique datasets. With this tool, the accuracy and generalizability of the investigation's findings improve as the number of unique datasets approaches infinity.

Model accuracy as a function of t/t ratio for one dataset:



Imagine what we could learn about the t/t ratio if we rerun this optimization algorithm with one million more datasets.

Future Development

Forthcoming developments:

- Graphical User Interface (GUI)

Questions for consideration:

- In what other manners may distributional characteristics of predictors vary between outcome groups?
- What are alternative ways to conceptualize the simulation of noise?

References & Author

Kim, B. S., Kang, B. G., Choi, S. H., & Kim, T. G. (2017). Data modeling in the big data era: Case study of a greenhouse control system. *Simulation: Transactions of the Society for Modeling and Simulation International, 93*(7), 579-594.

Schulz, A., Zöller, D., Nickels, S., Beutel, M. E., Blettner, M., Wild, P. S., & Binder, H. (2017). Simulation of complex data structures for planning of studies with focus on biomarker comparison. *BMC Medical Research Methodology, 17*, 90.

van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modeling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology, 14*, 137.

About Matt Myers

SCAN ME