# Fully Synthetic Neuroimaging Data for Replication and Exploration

Kenneth I. Vaden Jr., Ph.D.[1], Mulugeta Gebregziabher, Ph.D.[2], Mark A. Eckert, Ph.D.[1]

1. Hearing Research Program, Department of Otolaryngology-Head and Neck Surgery, Medical University of South Carolina.
2. Division of Biostatistics and Epidemiology, Department of Public Health Sciences, Medical University of South Carolina.

## Introduction

Fully synthetic neuroimaging datasets can facilitate replication and data exploration, while limiting privacy concerns and obstacles to data sharing and open science initiatives.

Data synthesis has increasing application as a discovery and educational resource, as well as when there are concerns about the risk of re-identification [Bellovin et al., 2019].

Fully synthetic data, when properly constructed, approximate observed data and statistical effects.

The current validation study was theoretically grounded in multiple imputation principles [Rubin, 1987; Rubin, 1993] and guided by multiple imputation findings with missing fMRI data [Vaden et al., 2012].

## Methods

**Participants:** Multi-site neuroimaging data for N = 264 children (107 F; age range = 6.39 to 12.85 years) from the Dyslexia Data Consortium [www.dyslexiadata.org; Eckert et al., 2016].

**Predictors:** Age, sex, Verbal IQ (VIQ) [Wechsler, 1999; 2004], intracranial volume (ICV), and study site.

**MRI Data:** T1-weighted images; denoised [Manjón et al., 2010]; bias-corrected, spatially normalized [DARTEL; Ashburner, 2007], modulated gray matter probability maps produced with SPM12 for standard voxel-based morphometry [Eckert et al., 2016].

**Analysis Strategy:** Observed data were transformed into multiple fully synthetic datasets (*m* simulants) through iterative random substitutions, then group statistics were performed for each simulant and averaged together to form a statistic estimate.

1. Predictor tables were synthesized (*m* = 10 versions).
2. Gray matter (GM) data were synthesized for simulated cases.
3. Synthetic t-score maps were averaged and optimally smoothed (0 to 2 mm FWHM maximized correlation with observed maps).

**Evaluating Synthetic Data:**

- **Efficiency:** synthetic data variability ÷ observed data variability.
- **Bias:** difference in synthetic — observed t-scores.

## Data Synthesis & Simulations

**A) Setup: Copy Dataset**

| | Age | Sex | VIQ | ICV | Site |
|---|---|---|---|---|---|
| original copy | 12 | M | 94 | 1204 | 1 |
| | 11 | F | 114 | 1155 | 2 |
| | 10 | F | 103 | 1040 | 3 |
| | 13 | M | 111 | 1290 | 4 |
| | . | . | . | . | . |
| | . | . | . | . | . |
| second copy | 12 | M | 94 | 1204 | 1 |
| | 11 | F | 114 | 1155 | 2 |
| | 10 | F | 103 | 1040 | 3 |
| | 13 | M | 111 | 1290 | 4 |
| | . | . | . | . | . |
| | . | . | . | . | . |

**B) Iteration 1: Removal**

| | Age | Sex | VIQ | ICV | Site |
|---|---|---|---|---|---|
| original copy | 12 | M | 94 | 1204 | 1 |
| | 11 | F | 114 | 1155 | 2 |
| | 10 | F | 103 | 1040 | 3 |
| | 13 | M | 111 | 1290 | 4 |
| | . | . | . | . | . |
| | . | . | . | . | . |
| second copy | 12 | ■ | 94 | 1204 | ■ |
| | ■ | F | ■ | 1155 | ■ |
| | ■ | ■ | 103 | ■ | 3 |
| | 13 | M | ■ | 1290 | ■ |
| | . | . | . | . | . |
| | . | . | . | . | . |

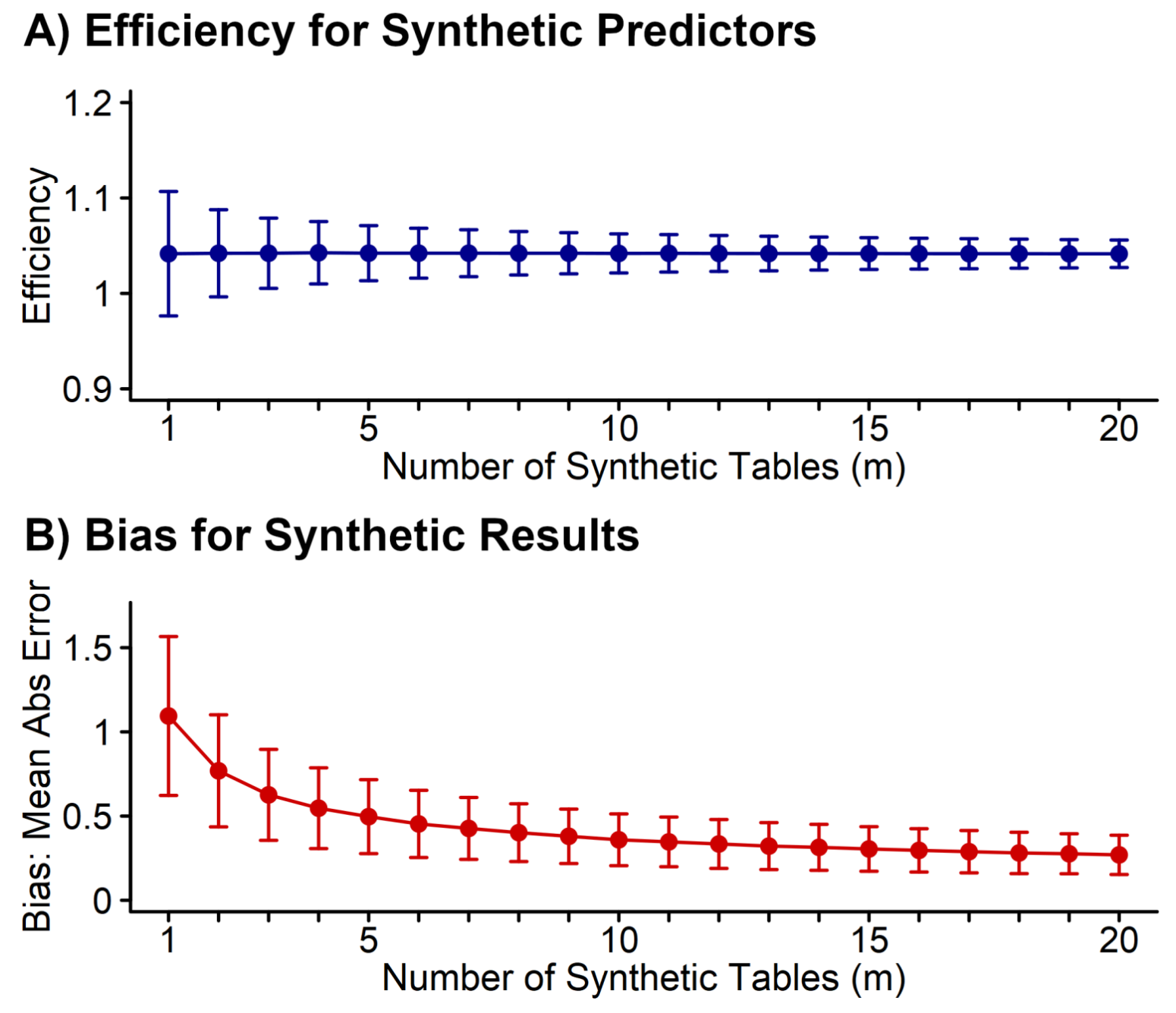
**C) Iteration 1: Imputation**

| | Age | Sex | VIQ | ICV | Site |
|---|---|---|---|---|---|
| original copy | 12 | M | 94 | 1204 | 1 |
| | 11 | F | 114 | 1155 | 2 |
| | 10 | F | 103 | 1040 | 3 |
| | 13 | M | 111 | 1290 | 4 |
| | . | . | . | . | . |
| | . | . | . | . | . |
| second copy | 12 | **F** | 94 | 1204 | **3** |
| | **12** | F | **111** | 1155 | **1** |
| | **9** | **F** | 103 | **1039** | 3 |
| | 13 | M | **108** | 1290 | **4** |
| | . | . | . | . | . |
| | . | . | . | . | . |

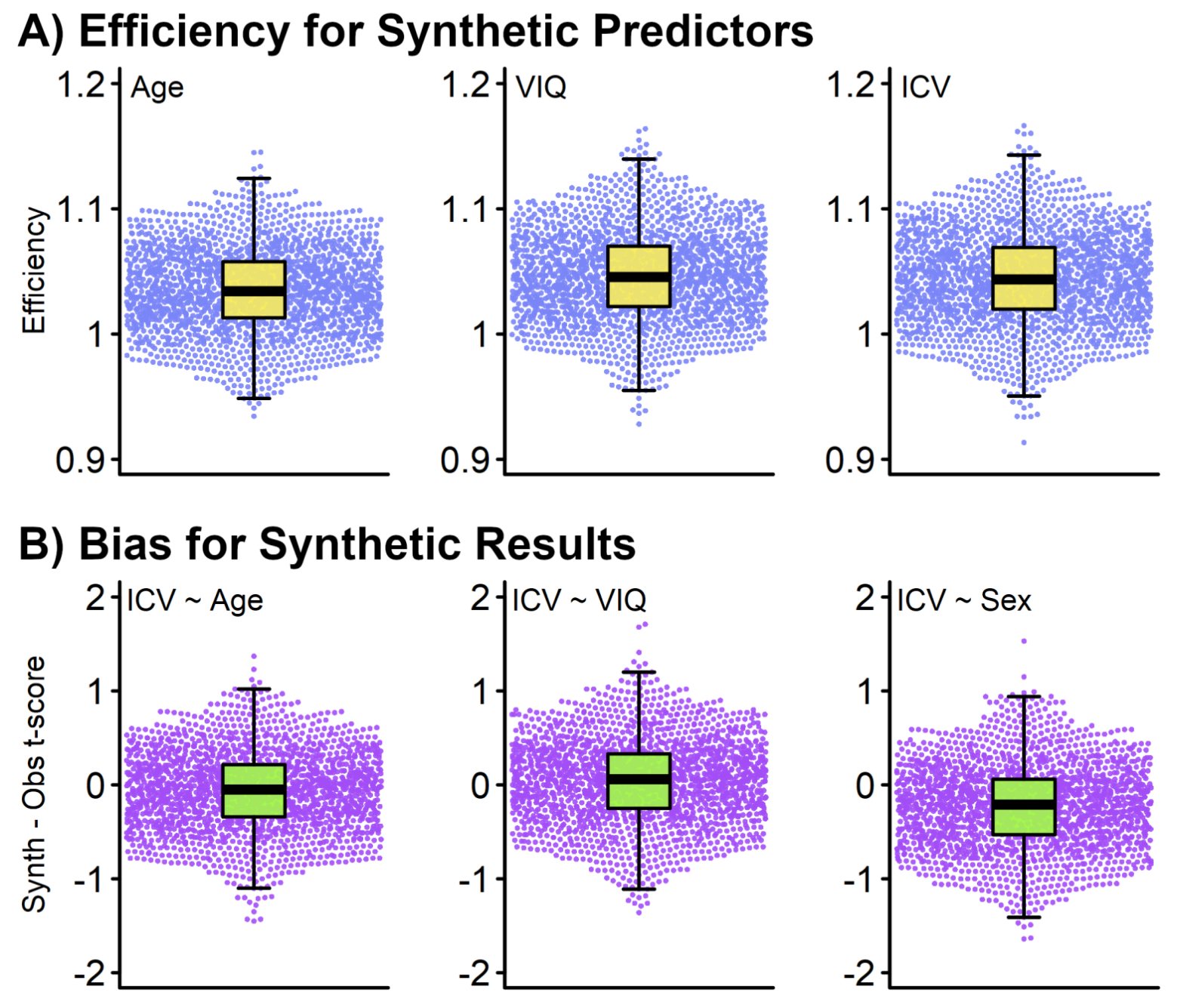
Observed data was iteratively removed, then substituted using an imputation model to replace missing values. [R-package: mice]

## Results: Fully Synthetic Data

### Synthetic Predictor Table Efficiency and Bias

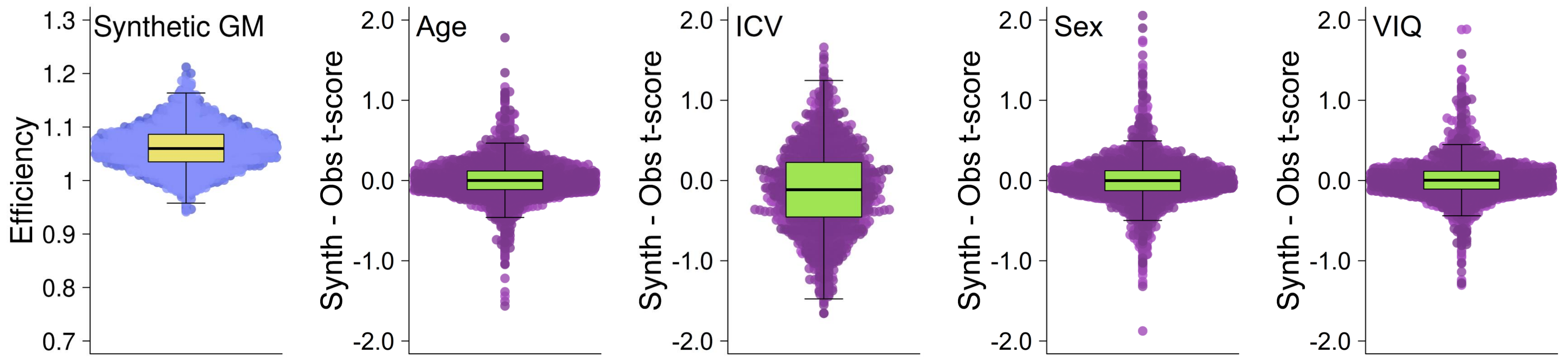


Simulation results also show that more simulants (*m*) increase the reliability of results, based on more stable efficiency (A) and lower bias (B).

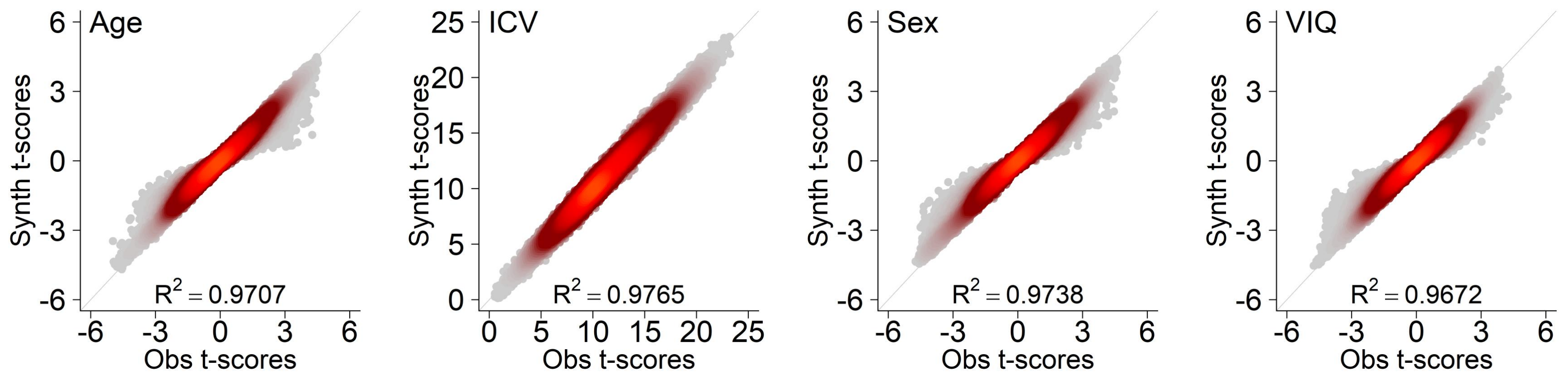


Efficiency and bias were averaged across *m* = 10 synthetic tables for 2,000 simulations (points). Efficiency scores close to 1 (A) limited statistical bias (B).

### Synthetic Gray Matter Efficiency and Bias



Mean efficiency = 1.05 for synthetic gray matter (GM) values (yellow boxplot, blue points), which were more variable than observed GM. Bias for statistic maps (green boxplots, purple points) was small and centered around zero. The larger ICV bias reflected larger effect sizes.

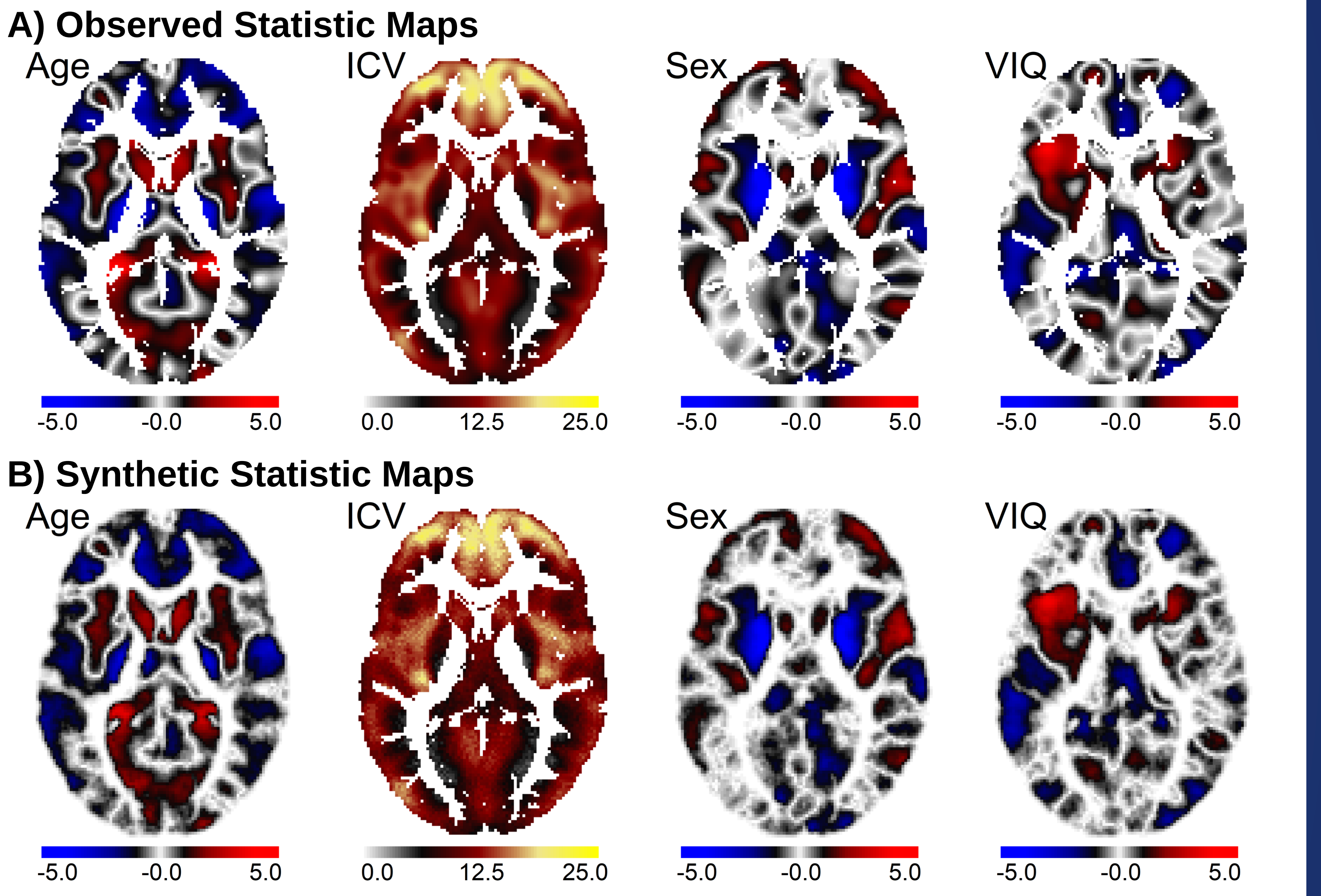


Gray matter associations with age, ICV, sex, and VIQ plotted in density scatterplots for voxel-level t-scores from each observed map (Obs) and smoothed average synthetic map (Synth).

## Results: Synthetic Statistic Maps

Smoothing average synthetic t-score maps (FWHM = 2 mm; except ICV) limited bias. Mean ± SD bias for age = 0 ± 0.25; ICV = -0.11 ± 0.51; VIQ = 0.01 ± 0.24; and sex = 0 ± 0.27.

R-square ≥ 0.97 across voxels in the observed and synthetic *t*-score maps, such that the results were nearly identical.



Comparison of observed (A) and synthetic (B) t-scores for a representative axial slice.

## Conclusions

Fully synthetic data can be used to accurately replicate results from real neuroimaging data.

The limitations of multiple imputation used for data synthesis must be considered (e.g., sample size, model compatibility).

Fully synthetic data has potential to enhance scientific integrity, discovery, and education.

**References.**

Ashburner J (2007): A fast diffeomorphic image registration algorithm. NeuroImage 38:95–113.
Bellovin SM, Dutta PK, Reitinger N (2019): Privacy and synthetic datasets. Stan Tech L Rev 1:1–51.
Eckert MA, Berninger VW, Vaden KI, Gebregziabher M, Tsu L, Dyslexia Data Consortium (2016): Gray Matter Features of Reading Disability: A Combined Meta-Analytic and Direct Analysis Approach. eNeuro 3:1–15.
Manjón J V., Coupé P, Martí-Bonmatí L, Collins DL, Robles M (2010): Adaptive non-local means denoising of MR images with spatially varying noise levels. J Magn Reson Imaging 31:192–203.
Rubin DB (1987): Multiple imputation for nonresponse in surveys. New York, NY: John Wiley & Sons. Vol. 26.
Rubin DB (1993): Statistical disclosure limitation. J Off Stat 9:461–468.
Vaden KI, Gebregziabher M, Kuchinsky SE, Eckert MA (2012): Multiple imputation of missing fMRI data in whole brain analysis. NeuroImage 60:1843–55.
Van Buuren S, Oudshoorn K (2011): MICE: Multivariate imputation by chained equations in R. J Stat Softw 45:1–67.
Wechsler D (1999): Wechsler Abbreviated Scale of Intelligence (WASI). San Antonio, TX: The Psychological Corporation.
Wechsler D (2004): The Wechsler Intelligence Scale for Children (WASI-IV). London: Pearson Assessment.

**Acknowledgements.** This work was supported (in part) by the National Institutes of Health (NIH) / Eunice Kennedy Shriver National Institute of Child Health and Human Development (R01 HD 069374) and was conducted in a facility constructed with support from Research Facilities Improvement Program (C06 RR 014516) from the NIH / National Center for Research Resources. Please see www.dyslexiadata.org for more information about and contributors to the Dyslexia Data Consortium.