NeuroImage 223 (2020) 117284

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Fully synthetic neuroimaging data for replication and exploration

Kenneth I. Vaden Jr.[a,*], Mulugeta Gebregziabher[b], Dyslexia Data Consortium, Mark A. Eckert[a,*]

[a] Department of Otolaryngology-Head and Neck Surgery, Medical University of South Carolina, 135 Rutledge Avenue, MSC 550, Charleston, SC, Unites States
[b] Division of Biostatistics and Epidemiology, Medical University of South Carolina, Unites States

ARTICLE INFO

Keywords:
MRI
Synthesis
Data sharing
Multiple imputation
Neuroimaging methods
Open science

ABSTRACT

Scientific transparency, data exploration, and education are advanced through data sharing. However, risk for disclosure of personal information and institutional data sharing regulations can impede human subject/patient data sharing and thus limit open science initiatives. Sharing fully synthetic data is an alternative when it is not possible to share real or observed data. Here we describe a data sharing approach that borrows principles and methods from multiple imputation to replace observed values with synthetic values, thereby creating a fully synthetic neuroimaging dataset that accurately represents the covariance structure of the observed dataset. Predictor tables composed of demographic, site, behavioral and total intracranial volume (ICV) variables from 264 pediatric cases were used to create synthetic predictor tables, which were then used to synthesize gray matter images derived from T1-weighted data. The synthetic predictor tables demonstrated pooled variance and statistical estimates that closely approximated the observed data, as reflected in measures of efficiency and statistical bias. Similarly, the synthetic gray matter data accurately represented the variance and voxel-level associations with predictor variables (age, sex, verbal IQ, and ICV). The magnitude and spatial distribution of gray matter effects in the observed imaging data were replicated in the pooled results from the synthetic datasets. This approach for generating fully synthetic neuroimaging data has widespread potential for data sharing, including replication, new discovery, and education. Fully synthetic neuroimaging datasets can enable data-sharing because it accurately represents patterns of variance in the original data, while diminishing the risk of privacy disclosures that can accompany neuroimaging data sharing.

1. Introduction

Data sharing is a key component of open science initiatives to enhance the integrity, impact, and pace of scientific discovery (Gorgolewski and Poldrack, 2016; Nichols et al., 2017), including through the secondary analysis of existing neuroimaging datasets (Brakewood and Poldrack, 2013; Poline et al., 2012). Sharing neuroimaging data is relatively easy when informed consent to share data is obtained (White et al., 2020), and software can be used to reduce privacy risks by de-identifying image datasets (Song et al., 2015). However, the risks of re-identification should be considered given the type and breadth of shared data (Dankar et al., 2012). For example, the inclusion of complementary data or unique populations increases the risk for re-identification, particularly when integrated with other datasets (e.g., in genetic studies: Gymrek et al., 2013; Homer et al., 2008). Privacy risks also differ across datasets with respect to the consequences of re-identification (e.g., stigmatized conditions or populations) and duration of risk (e.g., pediatric data). These risks may be higher for neuroimaging data collected through electronic health record databases where there can also be significant basic and clinical science value. Data managers, data providers, and data recipients must consider the costs and benefits of sharing de-identified datasets in the context of a shifting regulatory landscape for sharing information and biospecimens (Bledsoe et al., 2018), particularly with technological developments that may make it easier to re-identify data (Abramian and Eklund, 2019; Bellovin et al., 2019). Here we describe a new approach to synthesize neuroimaging, demographic, and behavioral data to: (1) facilitate replication for demonstrating the integrity of neuroimaging findings; (2) further limit the risk for re-identification; and (3) advance scientific discovery and education. The current approach complements neuroimaging data-sharing methods that typically balance data transparency and privacy risk (White et al., 2020).

Simulated brain images have been generated to evaluate image processing methods (Cocosco et al., 1997; He et al., 2015; Yang et al., 2015) and enhance computer-aided medical diagnoses (Castro et al., 2015), where realistic individual-level brain images were generated to train machine learning algorithms to make predictions or diagnoses based on images from individual subjects or patients. Influenced in part by concerns about false positive and false negative neuroimaging findings (Hong et al., 2019; Wager et al., 2009), an alternative examined here

* Corresponding authors.
E-mail addresses: vaden@musc.edu (K.I. Vaden Jr.), eckert@musc.edu (M.A. Eckert).

https://doi.org/10.1016/j.neuroimage.2020.117284
Received 27 February 2020; Received in revised form 12 August 2020; Accepted 16 August 2020
Available online 20 August 2020

Table 1
Study participant characteristics, MRI scanner, and T1-weighted image parameters.

Site	Sex	Age (M±SD)	Scanner	Dimensions (mm)	Slice Thickness (mm)	TR (ms)	TE (ms)	Flip Angle (deg)
1	30 F, 26 M	10.0±1.5	Siemens 1.5T	256×256×160	1.60	25	4.6	30
2	8 F, 11 M	9.0±1.5	Siemens 3T	176×240×256	0.90	2250	3.96	9
3	2 F, 8 M	9.8±1.7	Siemens 3T	128×256×256	1.33	6.6	2.9	8
4a	1 F, 8 M	11.8±0.6	GE 1.5T	124×256×256	1.20	11.1	2.2	25
4b	6 F, 13 M	11.0±1.0	GE 1.5T	124×256×256	1.40	11.1	2.2	25
5	14 F, 15 M	9.5±1.7	Siemens 3T	160×256×256	1.00	1600	3.37	15
6	3 F, 9 M	7.6±0.3	Philips 1.5T	170×256×256	1.00	8.02	3.69	7
7	16 F, 24 M	10.0±1.4	GE 1.5T	181×217×181	1.00	6	63	NA*
8	27 F, 43 M	8.8±1.1	Siemens 1.5T	160×256×256	1.00	2000	3.65	8

*Note: Flip angle was not reported for one study.

is to design the synthesis of data such that synthetic data accurately recapitulate group-level findings from real data.

Multiple imputation is one approach that can be used to generate synthetic data that accurately represents group-level results (Rubin, 1996, 1993). Here we describe a method for creating multiple "fully synthetic" datasets that closely approximate the observed data distribution (e.g., mean and variance structure) but include only artificial values and none of the observed cases (Loong and Rubin, 2017; Rubin, 1993). These fully synthetic datasets include multiple simulated versions of an observed dataset that together produce group-level statistical results closely approximating the "true" observed results.

Multiple imputation is a principled statistical approach that considers the distributions and co-variances of observed variables to produce multiple versions of a dataset that each substitute missing data with plausible simulated values (Rubin, 1987; Schafer, 1999). Any statistic of interest (e.g., mean, variance, t-score) is calculated separately for each simulated dataset, then averaged to form a point estimate that represents the value if the dataset contained no missing values. Averaging multiple "versions" of a group statistic effectively limits the influence of potential extreme simulated values in the imputed datasets. Previous findings have shown that multiple imputation can be effective in dealing with missingness in small and large datasets (Barnes et al., 2006), as well as mass-univariate neuroimaging datasets (Vaden et al., 2012).

We demonstrate that an imputation-based fully synthetic data approach can preserve the statistical properties of an observed neuroimaging dataset, which supports the replication of results and potential exploration of novel relationships with reduced risk of data-leakage and privacy breaches compared to de-identified data (Bellovin et al., 2019). This data synthesis approach, which was inspired by methods developed to protect the privacy of census/survey respondents (Nowok et al., 2016; Rubin, 1993), involves creating missing data by iteratively removing a proportion of data and then estimating the "missing" values using predictors (e.g., total brain volume for predicting missing gray matter data). That is, an imputation model was used to generate plausible synthetic values that reflected the distributions and variance structure in the observed data. After all the observed data were substituted with imputed values to produce multiple fully synthetic datasets, statistical tests were performed on each synthetic dataset then averaged together (i.e., pooled). Pooling statistical results across multiple versions of a simulated dataset can limit statistical bias from extreme simulated data points and allow valid inferences in the context of multiple imputation (Raghunathan et al., 2003).

The fully synthesized datasets were evaluated based on how well these approximated the variance and associations from the observed neuroimaging dataset. Specifically, participant characteristics and gray matter image data from a pediatric sample were used to validate the multiple imputation approach for generating fully synthetic data. Synthetic data replicated results from the observed data when variables of interest were included in both the imputation model and the analysis model. Research data were made available that include example code and data (Mendeley Data: Vaden et al., 2020), so readers can adapt these methods for their own research and data sharing purposes.

2. Materials and methods

2.1. Participants

The current study included retrospective, multi-site demographic and neuroimaging data selected from the Dyslexia Data Consortium (www.dyslexiadata.org). These data were received with Institutional Review Board (IRB) approval for sharing de-identified data from the contributing institution and with approval from the Medical University of South Carolina (MUSC) IRB to receive de-identified data.

Participants for this project were selected based on available T1-weighted images and included 264 children (107 females and 157 males; M ± SD for age = 9.55 ± 1.59 years; age range = 6.39 to 12.85 years) that were studied in Eckert et al. (2016). One motivation for this study was that some of the contributors of this data had IRB and institutional approvals to share data for a multi-site project at MUSC, but not to share the data more widely. Table 1 shows summary information for each site, including scanner and T1-weighted image parameters. These data provided an appropriate sample of convenience to assess with the synthetic data approach described here. Verbal Comprehension (VIQ), as measured by the Wechsler Intelligence Scales for Children (Wechsler, 2004) and the Wechsler Abbreviated Scales of Intelligence (Wechsler, 1999), was included as a predictor variable to inform the creation of synthetic data. A focus on VIQ was motivated by previous evidence that VIQ, and its shared variance with total gray matter volume, statistically explained a reading disability association with gray matter in a previous study (Eckert et al., 2016).

2.2. Image acquisition and preprocessing

Gray matter image data were used to assess the data synthesis approach because of their widespread use in voxel-based morphometry studies. T1-weighted structural images were preprocessed using MATLAB and SPM12 software for standard voxel-based morphometry, as in our previous brain morphology studies (Eckert et al., 2017, 2016). Details are provided for each preprocessing step to enhance the likelihood of replicability (Poline et al., 2012). First, images were bias-corrected and denoised using the adaptive non-local means algorithm, which estimates then removes spatially varying noise from T1-weighted images (Manjón et al., 2010). Each brain image was then rigidly aligned to the MNI template using the SPM12 co-register function based on the default, normalized mutual information objective function. The SPM12 segmentation function was then used to create gray matter, white matter, and cerebrospinal fluid probability maps, with default selections for bias FWHM (60 mm), bias regularization (0.001), MRF parameter (1) and clean up (light clean). Total intracranial volume (ICV) for each participant was calculated by summing voxel values across the three probability maps. The native space gray matter images were spatially transformed into a study-specific neuroanatomical space using the SPM12 diffeomorphic normalization procedure with default parameter selections, including linear elastic regularization with 6 outer warping iterations (each with 3 inner iterations with decreasing regularization

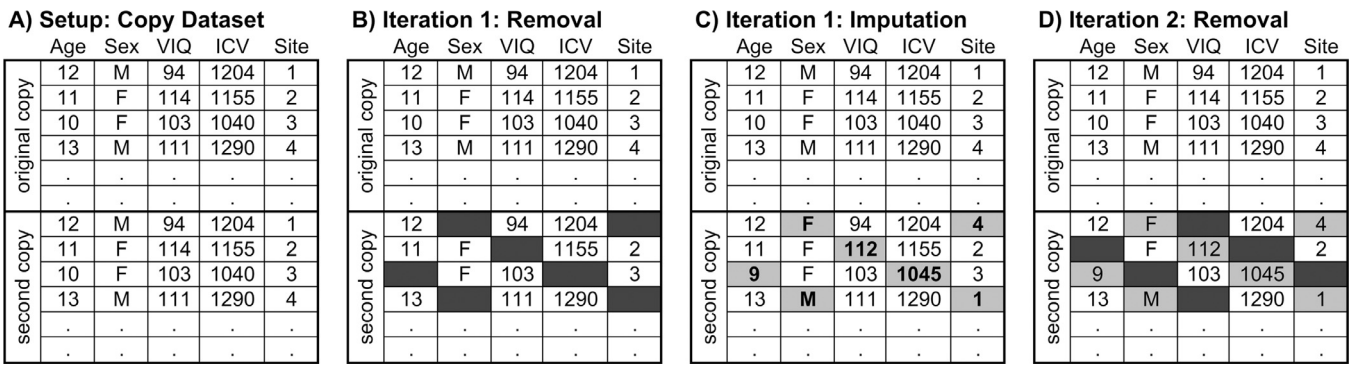

Fig. 1. Predictor table synthesis overview. (A) Two copies of the original data were attached to set up the data matrix. (B) Observed values were removed in randomly ordered but fixed intervals from the second copy during each iteration (dark gray and empty cells). (C) Next, the cells with missing data were filled with imputed values. (D) Observed values were removed and replaced with simulated values at fixed intervals in the second copy of the data, such that second copy of data was completely substituted with simulated values over the course of a few iterations. The fully synthetic dataset was then separated from the unaltered original copy. The iterative procedure was repeated to create $m = 10$ independent and fully synthetic datasets for the current study. R code for performing this procedure is available online (Mendeley Data: Vaden et al., 2020).

parameters), where the average group image from the preceding step served as the normalization target for the next step (DARTEL; Ashburner, 2007). Spatially normalized gray matter maps were modulated using the normalization deformation parameters, which created gray matter volume images or images where each gray matter voxel was weighted by the extent of volumetric displacement to that voxel during normalization. Each modulated gray matter probability map was smoothed with a Gaussian kernel (FWHM = 8 mm), commonly used to limit false positive results and better approximate normally distributed data.

2.3. General framework for creating fully synthetic data in two steps

The current data synthesis approach focuses on replicating group-level statistical results rather than creating realistic individual subject data. Synthetic data were produced using a two-step process. The first step was to generate multiple synthetic predictor tables. The second step was to generate multiple sets of synthetic gray matter images, one for each synthetic predictor table. For both the predictor tables and images, data was iteratively substituted by removing and then imputing values until a completely new synthetic dataset was created. This approach allowed simulated data to inform subsequent imputations for each iteration, such that the average associations between variables in the synthetic data would accurately reflect associations in the observed data.

Each synthetic dataset was created by attaching two copies of the observed data together within one large data matrix (Fig. 1A). The first copy of the data was never perturbed, which contributed to preserving the variances and associations between the observed and synthetic data. The second copy underwent iterative data substitution (Fig. 1B to 1D). During each iteration, the second copy of the data was increasingly composed of imputed values. By the end of the last iteration, the second copy of the data was fully composed of synthetic values. Again, the first copy of the original data preserved the real covariance structure to inform the imputation model, while the second copy became increasingly synthetic. Two copies of the dataset also meant that 10% data removal in the second copy only resulted in 5% artificial missingness for the imputation model, where imputation provides valid inference (Vaden et al., 2012). We performed simulation tests to confirm that synthetic results more accurately represented the observed results when less data was substituted during each iteration of the imputation-based data generation process (Supplementary Fig. S1).

The iterative data replacement procedure was repeated to create multiple (m) versions of the synthetic dataset, borrowing from the methods and principles of multiple imputation. Parameter estimates (e.g., average, standard deviation, t-scores) were calculated separately for each m version of the synthetic dataset, and then averaged to accurately represent the observed parameters and results (Rubin, 1996, 1993, 1987). Averaging parameter estimates from multiply imputed data is known to create reliable point estimates for group-level results, and the number of fully synthetic datasets (m) needed for valid synthetic results is considered later. In summary, the current approach generated m synthetic versions of the predictor table and image data to ensure that group-level associations were preserved while eliminating data from real cases, and theoretically re-identifiable unique combinations of data (El Emam et al., 2011).

2.4. Fully synthetic predictor tables (step one)

Predictor tables were created that included the following variables: age, ICV, sex, VIQ, and the research site label for each participant in the multi-site dataset. The R-package 'mice' (version 3.6.0) (Van Buuren and Groothuis-Oudshoorn, 2011) was used to create an imputation model based on observed and simulated values in the dataset to inform the generation of plausible replacement values. Multiple versions of the fully synthetic predictor tables were created ($m = 10$), which each contained unique simulated cases (i.e. "simulants") that could not be linked to any observed case. Because MICE uses distributional information to replace values, individual values from the original dataset that are unusual and/or identifiable are unlikely to be exactly replicated in the simulants. This replacement approach differs from shuffling or exchanging observed values to simulate cases, which can propagate unusual values that are potentially unique identifiers for some datasets (e.g., age > 89 years). The example R code (Mendeley Data: Vaden et al., 2020) includes a function to verify that no synthetic cases match an observed case, which was also confirmed for the current study.

2.5. Evaluation of synthetic predictor tables

Numerous measures have been used to evaluate the quality of simulated data for multiply imputed datasets. For example, Stuart et al. (2009) suggest that imputed variables should have less than twice and more than half the variance of the observed data, as well as an absolute difference in the simulated and observed means less than two standard deviations. Kolmogorov-Smirnov tests can also be used to detect significant distribution differences between observed values and imputed values (Abayomi et al., 2008). The current study used variance ratio measures, t-score differences, covariance tests, and correlation-based comparisons of each observed and synthetic variable to assess the fully synthetic predictor tables and neuroimaging data.

First, synthetic variance was characterized to determine how accurately it represented observed variance on average, for each variable or voxel in the synthetic datasets. Efficiency was defined as the average ratio of synthetic variance to observed variance, for each variable in the synthetic data which included participant age, ICV, sex, and VIQ. Efficiency is ideally equal to one, such that the variance for a synthetic variable matches the variance for the observed variable. Efficiency pooled across many simulations that differs substantially from one (e.g., efficiency < 0.05 or efficiency > 2) can indicate bias in the variance estimate used to simulate values. Efficiency < 1 indicates that a synthetic variable had lower variance than the observed variable, and efficiency > 1 indicates that a synthetic variable had higher variance than the observed variable.

Accuracy of the synthetic statistical test results (e.g. t-scores) was determined using a measure of bias, or the absolute difference between an observed t-score and the average synthetic t-score. Bias was defined as the absolute difference between an observed t-score and the average synthetic t-score based on each of the synthetic predictor tables. Specifically, bias was examined for the synthetic predictor tables by performing separate GLM-based regression analyses that each specified ICV as the outcome variable and participant age, sex, or VIQ as the explanatory variable, and then contrasting the pooled synthetic t-score to the observed t-score from each regression test. The absolute difference between the pooled synthetic t-score minus the observed t-score represented the magnitude of estimation error for each statistical test performed with the synthetic and observed datasets.

The quality of the synthetic data was judged based on preservation of covariance structure relative to the observed predictor table. Because the predictor tables included a mixture of continuous and non-ordinal categorical data (e.g., sex, research site), mutual information was calculated for each pair of variables using Mixed-Pair Mutual Information Estimators (R-package: "mpmi", version 0.43). Pairwise mutual information estimates were organized into mutual information matrices (MIM), similar to a covariance matrix, then a correlation test was used to quantify the association between synthetic and observed MI estimates. The resultant correlation coefficient quantified the degree to which observed pairwise associations were accurately represented in the synthetic datasets.

2.6. Number of synthetic datasets

Simulation tests were performed to evaluate the number of synthetic predictor tables (m) needed to accurately represent variance and associations from an observed dataset, and to guide the number of fully synthetic datasets generated for the current study. Researchers often use ten or more imputations for statistical analyses with missing data, although the number of imputations necessary to limit inference error depends on effect size, sample size, and the extent of missingness (Lu, 2017; Rubin, 1996; Schafer, 1999).

During each simulation test (2000 repetitions), data from the observed predictor table was replaced with imputed values to create 10 synthetic predictor tables. Consistent with multiple imputation principles, parameter estimates (e.g., variance, mutual information, and t-scores) were calculated separately for each synthetic predictor table, and then pooled estimates were calculated from $m = 1$ to 25 estimates, for each simulation. Efficiency, mutual information, and bias results were examined across simulations to determine how the number of synthetic datasets generated ($m = 1$ to 25) can affect the approximation of observed data. The Supplementary Materials include results from additional simulation tests to characterize how synthetic data efficiency and bias were related to sample size, percent data replacement, and number of synthetic datasets generated. The simulation test results and supplementary results provide a rationale for the $m = 10$ and iterative replacement of 10% data in the current study. These results may also provide guidance for researchers interested in generating novel synthetic datasets (Supplementary Figs. S1 and S2).

2.7. Fully synthetic neuroimaging data (step two)

Based on multiple imputation guidelines and results from the simulation described above, synthetic gray matter datasets were produced for each of the $m = 10$ synthetic predictor tables. Similar to the approach for generating synthetic predictor tables, the observed predictor table and gray matter values for a voxel (observed data) were temporarily attached to a copy of the synthetic predictor table. Next, synthetic gray matter values were produced for the "missing" gray matter image values using multiple imputation. Thus, the imputation model was again guided by associations between the observed predictors (e.g., age) and gray matter data when simulating data for each voxel.

Efficiency was calculated for each voxel as the ratio of synthetic variance to observed variance in the gray matter images, to characterize potential bias in the variance estimates from the imputation model. Mutual Information was again used to measure the associations between the mixed-type predictors (categorical and continuous) and gray matter volume for each voxel. A correlation test was then performed between the observed mutual information and synthetic mutual information, with the expectation that well-preserved associations would be reflected in strong, positive correlations across voxels. The associations between gray matter volume and predictor variables were used to quantify the validity, and thus usefulness, of the synthetic data.

Statistical parametric maps were produced to examine the spatial pattern of the observed and simulated results. Whole-brain GLM regression analyses were performed using SPM12 with gray matter image data as the outcome variable, and participant age, ICV, sex, and VIQ as explanatory variables. The resultant t-score maps were pooled across simulations ($m = 1$ to 10) for comparisons with the observed t-score maps. Each pooled synthetic t-map was tested to determine the optimal Gaussian smoothing kernel that would maximize its correlation with the observed t-score map across a test range of FWHM = [0, 0.5, …, 12 mm]. Smoothing the synthetic results was necessary because gray matter values in each voxel were synthesized independently of the adjacent voxels, which created spatially independent variance in the t-score maps. For the same reason, simulated gray matter images could not be smoothed prior to statistical analyses because reducing the spatial variance in synthetic subject images can potentially inflate group-level results. Bias was used to assess differences in the statistical result maps, by calculating the absolute difference between observed and synthetic t-score in each voxel.

In addition to characterizing efficiency for the simulated brain images and bias for the t-scores, analyses were performed to characterize agreement between observed and synthetic SPM results across small to large effect sizes. Each statistical map was thresholded with uncorrected cluster defining p-value thresholds (CDF $p = 0.05, 0.01, 0.001, 0.0001$] and non-stationarity correction in SPM12 (Hayasaka et al., 2004) was used to identify significant clusters based on extent with familywise error corrected $p = 0.05$. Hits were defined as significant voxels or clusters with at least 50% overlap, which were present in both the synthetic and observed statistical maps. False alarms were defined as significant voxels or clusters that were present in the synthetic but not the observed statistical map. Misses were defined by significant voxels or clusters that were present in the observed but not the synthetic statistical map. True positive rates were calculated as the number of hits divided by the total number of significant voxels or clusters in each observed statistical map. Hit rates were calculated as the total number of hits divided by number of hits and false alarms, to determine the proportion of significant results in the simulated t-score maps that correspond to the observed significant results.

2.8. Imputation model specification

The inclusion of a variable in the analyst model that is not included in the imputation model can have negative consequences on the inference of results from multiply imputed data (Meng, 1994; Rubin, 1996).

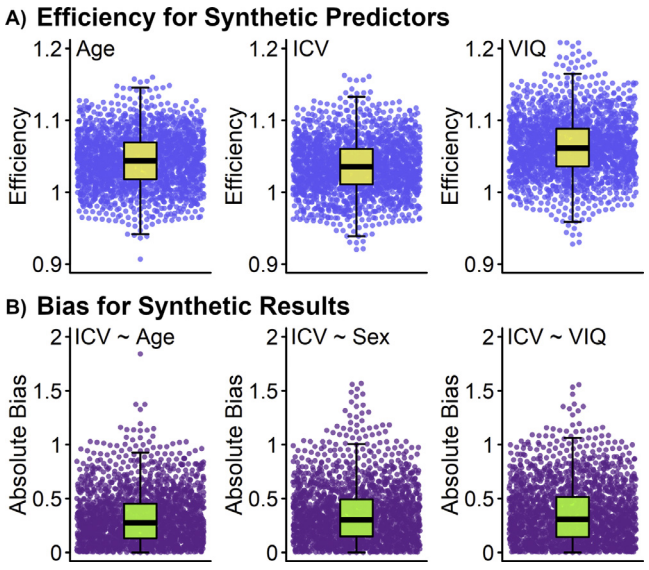

Fig. 2. Efficiency and bias were averaged across 10 synthetic predictor tables for each of the 2,000 simulations (plotted as points). The synthetic predictor tables showed efficiencies ≈ 1 and bias ≈ 0. **(A)** Efficiency quantiles for predictors are shown as boxplots. **(B)** Parameter estimate bias for t-scores from a model of ICV predicted by age, sex, or VIQ are displayed with boxplots.

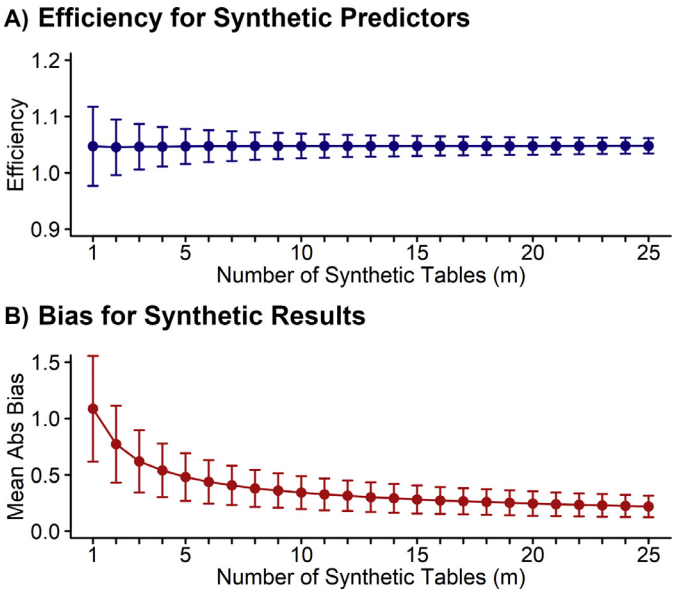

Fig. 3. The simulation results suggest that generating more synthetic predictor tables can increase reliability, based on more stable efficiency and lower bias. **(A)** Synthetic data had higher variability than the observed data, such that efficiency was consistently just above 1. Efficiency did not appear to decrease in relation to a larger number of synthetic tables, although efficiency was more stable based on smaller SD error bars. **(B)** Pooled statistical parameter estimates showed lower bias when results were averaged across a larger number of synthetic tables, based on the mean absolute difference between observed and synthetic t-scores. The SD error bars also suggest that bias was less variable with a larger number of synthetic predictor tables. Abs: Absolute.

We examined the extent to which the current approach for data synthesis using multiple imputation was also sensitive to this limitation. Specifically, we examined if Sex-VIQ interactions in the observed results would be misrepresented in the synthetic results. A false negative bias was expected for an interaction term that was excluded from the imputation model used to generate synthetic values. Results presented in the Supplementary Materials show that this well-known modeling consideration of multiple imputation for missing data also applies to fully synthetic data (Supplementary Figs. S3 and S4).

2.9. Data and code availability statement

As noted above, there were institutional limitations on data-sharing for a subset of the data used for this study. A research dataset that includes the synthetic predictor tables and fully synthetic neuroimaging datasets are available online (Mendeley Data: Vaden et al., 2020). The research dataset also includes commented MATLAB and R code to implement the current neuroimaging data synthesis methods with a small example dataset. The example data were selected from an earlier fMRI study (Kuchinsky et al., 2012) to demonstrate that the current method can be applied to other types of neuroimaging data. The example code can also be adapted to produce fully synthetic group-level datasets based on observed neuroimaging data from other sources.

3. Results

3.1. Synthetic predictor tables evaluation

The simulation results with pooled estimates from multiple synthetic predictor tables (m = 10) showed that the observed co-variance structure, variances, and associations were accurately represented by the fully synthetic predictor tables. Large correlation coefficients for the MIMs (mean $r = 0.93 \pm 0.01$; range = 0.89 to 0.96) suggested that the covariance structure of the synthetic data closely approximated the variance structure of the observed data. Fig. 2A shows that efficiency was typically just above one for each variable in the synthetic predictor tables (mean efficiency: age = 1.04 ± 0.04; ICV = 1.04 ± 0.04; VIQ = 1.06 ± 0.04). Bias was also limited across simulations (Fig. 2B; absolute bias for ICV predicted by: age = 0.32 ± 0.24; sex = 0.35 ± 0.26; VIQ = 0.36 ± 0.27). Together, the simulation results indicated that the fully synthetic predictor tables accurately represented the variances and associations within the observed data from which they were derived.

3.2. Number of synthetic predictor tables

The simulation results were also used to characterize the cumulative statistical benefit from pooling results from an increasing number of synthetic predictor tables. Fig. 3A shows that variance of the synthetic data was higher than the observed data regardless of how many tables were synthesized. However, the error bars in Fig. 3A indicate that efficiency was more reliable (less variable) across simulations when more synthetic predictor tables were generated. The bias results also suggest that pooled t-score estimates were more accurate and less variable across simulations based on a larger number of synthetic predictor tables (Fig. 3B). While higher m values reduce variability for efficiency and lower statistical bias, these benefits come with the computational burden of generating and analyzing many synthetic datasets. This is especially the case for relatively large samples, such as the current N = 264. Because each result stabilized and produced diminishing returns after 10 imputations, m = 10 fully synthetic datasets were generated for the current study. Supplementary Figs. S1 and S2 provide related information for a simulated range of smaller sample sizes.

3.3. Synthetic gray matter image evaluation

The synthetic and observed predictor tables were used to create synthetic images, as described in the Methods. Fig. 4 shows that (unsmoothed) synthetic image data were more variable than observed image data (mean efficiency = 1.07 ± 0.04). The observed covariance structure was well-approximated in the synthetic image data, based on mutual information between each predictor and gray matter values (mean MIM correlation $r = 0.96 \pm 0.06$). These results suggest that the

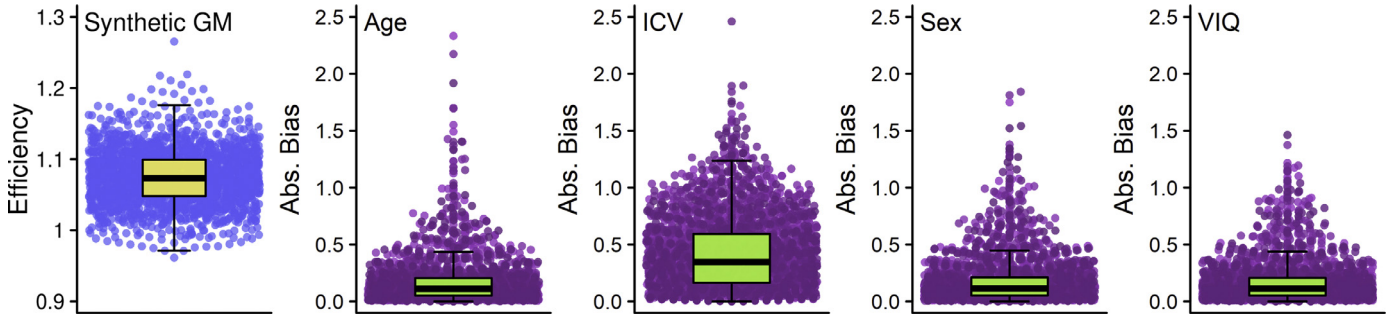

Fig. 4. Efficiency and bias were used to characterize how well the observed gray matter data and results were approximated by the fully synthetic dataset. The yellow boxplot on the left illustrates efficiency quartiles, which are overlaid on blue points that show the efficiency values from individual voxels. Efficiency was higher than one in most voxels, which means that simulated gray matter values were more variable than observed values. The green boxplots show bias quartiles overlaid on purple points for voxel-level bias estimates, calculated as the absolute difference in t-scores from the observed and synthetic statistical maps (Synth - Obs). Bias was relatively close zero. The larger magnitude for ICV bias appears to reflect the strong association between brain size and voxel-level grey matter volume.

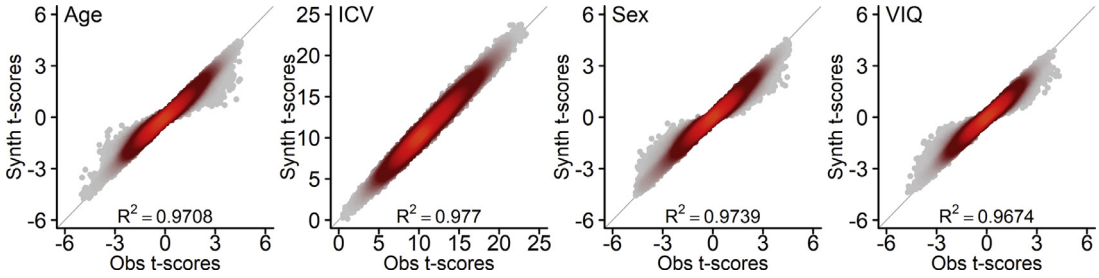

Fig. 5. A multiple regression model was used to test for gray matter associations with age, ICV, Sex, and VIQ. Density scatterplots show the concentration of voxel-level results for each observed and synthetic statistical map. The t-scores from the observed statistical maps (Obs) were well-approximated by t-scores in the synthetic statistical maps (Synth). The light gray points indicate the lowest density of overlapping points, dark red points indicate higher overlap density, and bright red points indicate the highest overlap density in each subplot.

fully synthetic dataset accurately represented gray matter image values and associations with predictors in the observed dataset.

Because synthetic image data were generated independently for each voxel with an imputation model naïve to spatial dependencies with neighboring voxels, synthetic image data and statistical maps were more variable than the observed images and results. As described in the methods, synthetic statistical maps were generated by performing multiple regression analyses and averaging $m = 10$ simulated versions of each result, then spatially smoothing the pooled result with an optimized Gaussian smoothing kernel. The best smoothing kernel size was determined for each pooled synthetic statistical map by the highest correlation between each observed statistical map and its synthetic counterpart. The pooled synthetic statistical maps for age, sex, and VIQ were optimally smoothed with a FWHM = 2 mm smoothing kernel, and ICV was smoothed with FWHM = 0 mm (i.e., not smoothed). Below, the pooled and smoothed synthetic results are referred to as: “synthetic statistical maps”.

The absolute differences between the observed and synthetic t-score maps were relatively small (M ± SD absolute bias for Age = 0.16 ± 0.19; ICV = 0.41 ± 0.31; VIQ = 0.16 ± 0.18; Sex = 0.17 ± 0.19). The limited voxel-based gray matter bias is consistent with efficiency close to one for the synthetic predictor tables and synthetic gray matter data (Fig. 4).

Consistent with the relatively low bias, the synthetic results accurately represented the observed results, based on correlations performed across voxels on the t-scores from the observed and synthetic statistical maps (Fig. 5). Strong correlations were obtained between the observed and synthetic statistical maps for age, ICV, sex, and VIQ, such that these were nearly identical ($R^2 \geq 0.97$). The relatively infrequent underestimation of an observed result (e.g., propeller shapes in the Fig. 5 Age effects) was consistent with efficiency > 1 for the synthetic age and synthetic gray matter image data. Fig. 6 displays an axial slice from each of the observed and synthetic statistical maps to show the spatial consistency. Together, these results indicate that the synthetic image data accurately reflected associations between gray matter image data and predictor variables in the observed data.

The whole-brain non-stationarity corrected cluster results from SPM12 also suggest that synthetic t-score maps accurately represented the observed results (Fig. 7). Nearly all cluster-level hit rates and true positive rates were equal to 100%, except for the true positive rate when the cluster defining threshold $p = 0.0001$. To better understand these cluster results, the voxel-level hit rates were examined and these hit-rate data demonstrated a similar pattern. The voxel level hit rates dipped from 82.7% to 39.4% as the cluster-defining p value was lowered from 0.05 to 0.0001, which when considered with the bias results shown in Fig. 5, demonstrate that the synthetic results had lower t-scores and this affected the cluster results when using a high p value threshold. True positive rates for the voxels ranged from 97.3% to 98.8%, suggesting that the significant clusters in synthetic statistical maps rarely extended beyond the spatial extent of the observed clusters. Together, these results show that fully synthetic neuroimaging results replicated the observed statistical maps, which is consistent with efficiencies near one and low bias estimates in non-thresholded t-score maps. Nonetheless, efficiencies > 1 can limit the correspondence of significant results when using conservative p value thresholds to define clusters in statistical maps.

4. Discussion

The results from this study provide validation for a data synthesis approach that is conceptually based on the principles of multiple imputation (Rubin, 1993). We demonstrated that synthetic predictor variables (age, sex, VIQ, ICV) and gray matter image data closely approximated the covariance structure, variance, and statistical estimates in the observed dataset. We envision that the type of synthetic data

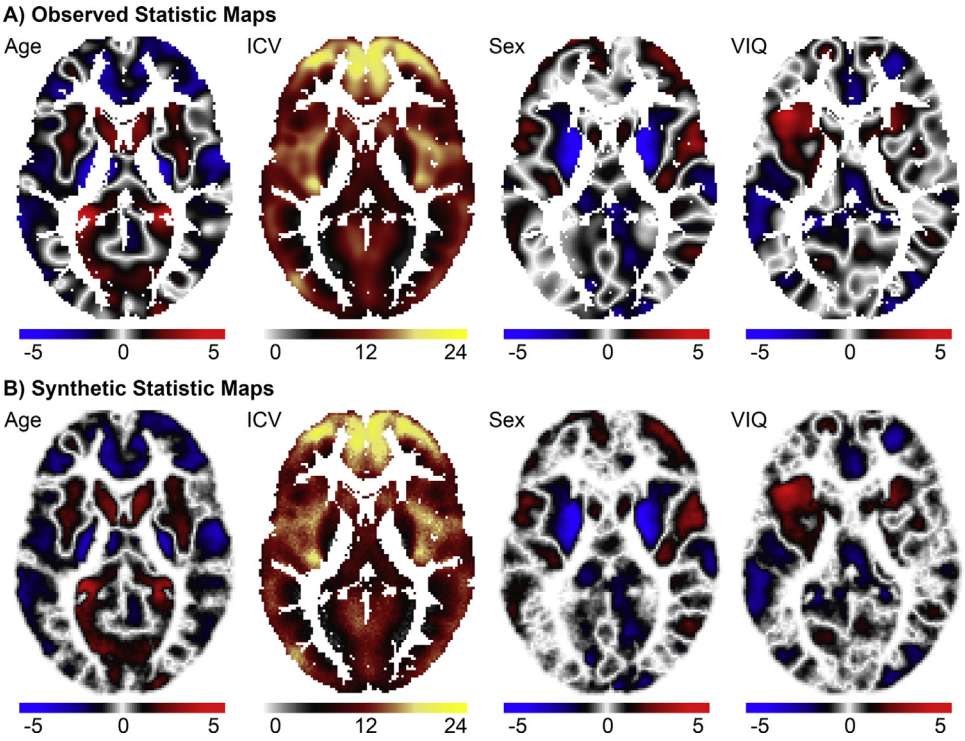

Fig. 6. An axial slice is shown for the (**A**) observed statistical maps and (**B**) synthetic statistical maps. The color scale for the t-scores is shown below each statistical map.

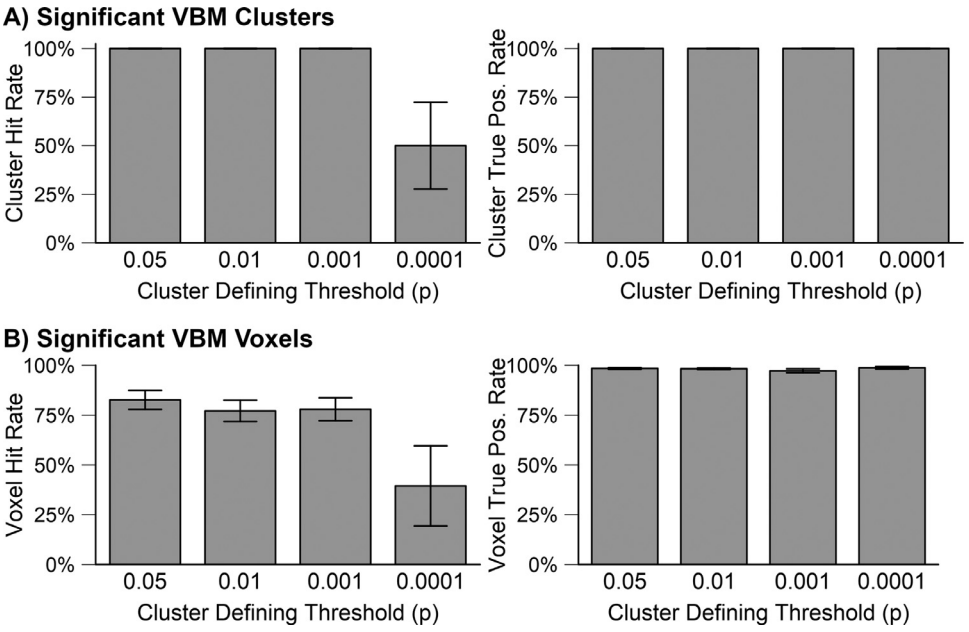

Fig. 7. (**A**) Comparing the significant clusters for synthetic t-score maps and observed t-score maps indicated that results were consistent, as demonstrated by hit rates = 100% at typical cluster defining thresholds. The cluster true positive (Pos) rates indicate that no significant clusters appeared in the synthetic maps that were not also present in the observed results. Significant cluster hit rates decreased as the cluster defining *p* value was lowered, which is consistent with efficiency values greater than one for the imaging data and a false negative bias. That is, the synthetic result clusters were smaller in extent than the observed result clusters. This interpretation is consistent with results shown in **B** where the hit rate for voxel-level results also decreased with each more conservative *p* value. Because true positive rates were close to one, the results also suggest that false positives were rarely present in the synthetic t-score maps. Each bar in the subplots represents the average ± SEM bars for the positive and negative signed contrasts for age, ICV, sex and VIQ.

generated in the current study could be used for data sharing initiatives, including for manuscript and grant application review, limited datasets of patient information in medical institution clinical databases, algorithm development, and for educational purposes. Users of this data synthesis approach should carefully consider the well-established guidelines and limitations for multiple imputation.

4.1. Fully synthetic and unidentifiable data

Our fully synthetic data approach is focused on characterizing group-level results in the observed data rather than creating individual subject data. One benefit of this approach is that it significantly limits the possibility that any single case in the observed data can be identified. While the current study included a large multi-site sample with low risk for privacy disclosures, de-identifying data can be problematic for studies where cases may be identified based on unique combinations of variables or unusual variable values (Dankar et al., 2012; El Emam et al., 2011). Across 2000 simulations with 10 synthetic predictor tables (264 cases each), no synthetic case in the current study had a combination of demographic, behavior, site, and ICV values that exactly matched one of the 264 observed cases. However, it is possible that identical combinations of variable values can appear in observed and synthetic predictor tables. For that reason, the code for generating synthetic data included a function to detect simulants that were identical to cases in the observed dataset. If an identical pattern of variable values is identified, the code generates another synthetic dataset to replace all simulants. This

function is included in the example code available with the research dataset (Mendeley Data: Vaden et al., 2020), and provides a safety check to identify and remove potentially duplicated patterns of observed data from the synthetic datasets.

4.2. Synthetic data creation and use recommendations

The extent to which synthetic data accurately replicate observed results appears to be sensitive to the proportion of "missing" data that are iteratively substituted with imputation during the data generation process. Here 10% of the values were removed and imputed during each iteration to synthesize predictor tables. Because the imputation model uses an intact copy of the original table, this represents 5% missingness for the imputation model. Fewer missing data can sometimes improve the quality of multiply imputed data, particularly for smaller sample sizes (Vaden et al., 2012). Our supplementary results show that replacing fewer values at each iteration during the generation of fully synthetic data produced better efficiency and lower bias, especially for smaller samples (Supplementary Fig. S1). Increasing the number of synthetic tables can further reduce bias for synthetic t-scores (Supplementary Fig. S2). These observations suggest that modifications to the current approach, such as substituting a smaller proportion of data using imputation or increasing the number of synthetic datasets, may be important for smaller sample sizes.

In addition to how well the synthetic data represent the observed data, the usefulness of the synthetic data will depend on the recipient's understanding of the synthetic and observed datasets. At a minimum, recipients of synthetic data will need access to the imputation model to evaluate the synthetic data and know what variables can be appropriately analyzed (Reiter, 2005). This is because the recipient's analysis model should include the same variables as the contributor's imputation model (Loong and Rubin, 2017). For example, an interaction tested with synthetic data can produce false positive and/or false negative results when the imputation model does not include that interaction (Tilling et al., 2016). Supplementary Figs. S3 and S4 demonstrate that synthetic gray matter results were statistically biased when the data synthesis model did not include a sex-VIQ interaction effect that was specified for the multiple regression model (Meng, 1994; Rubin, 1996).

Maximizing the imputation model or including all variables and potential interactions of interest in the predictor model during the generation of synthetic datasets could potentially remedy the imputer/analyst model consistency problem, but this may be computationally unwieldy depending on the number of higher-order interactions and dataset size. We recommend clear communication between data contributors and recipients to better understand the imputation model used to generate the synthetic dataset (Loong and Rubin, 2017), or the development of online data portals that allow users to specify an imputation model that can generate synthetic data tailored for their analysis plan. Allowing recipients to control data generation models could address challenges in the analysis of imputation-based synthetic data that can result from specification differences between the imputer model and the analyst model. We also recommend that recipients request that contributors replicate any novel findings with their observed data, when the recipients discover novel findings based on synthetic data. We return to this issue later in the context of false positive and false negative findings.

Contributors of synthetic data would also ideally provide statistics about the similarity of their synthetic and observed data. We used efficiency and bias in the current study, but other metrics could be important. For example, efficiency at each voxel could be provided so that recipients understand which brain regions exhibit variance that is most similar (or dissimilar) to the observed variance. Given that synthetic data generation is focused on providing accurate parameter estimates, images or maps of the correspondence between synthetic and observed results (e.g., Fig. 5) could also be provided to recipients. We note that the example code available from our research dataset (Mendeley Data: Vaden et al., 2020) can be used to provide this information about a fully synthetic dataset.

4.3. Limitations and cautionary guidance

The accuracy of the imputation approach for data synthesis depends on associations within the predictor table and image data. When the synthetic data have efficiency ≈ 1, then false negative and false positive effects in the synthetic data should be minimized. This is important in the context of caution about false positive findings in neuroimaging studies (Eklund et al., 2019; Greve and Fischl, 2018; Scarpazza et al., 2015). The examination of efficiency and bias in synthetic results could be useful to researchers who have concerns about the ability of other researchers to replicate their findings. In the current study, there was evidence of a false negative bias when there was increased variance in the synthetic data relative to the observed data (i.e., efficiency > 1). While the current study serves as a useful proof of concept, the efficiency results suggest that data synthesis could be enhanced with the addition of more sensitive predictor variables.

We also caution against extrapolating values well beyond the observed data range to artificially create greater variance in synthetic data. This could produce invalid inferences because the associations between variables may be non-linear for values beyond the observed range. Moreover, it seems likely based on the efficiency observations that this could also create false negative results.

The degree to which effects replicate between synthetic and observed datasets may also depend on the sample size. The current study involved a relatively large sample size ($N = 264$) for a structural neuroimaging dataset. This approach also appears effective for a relatively smaller fMRI dataset ($N = 36$) that was used for the example code. While this example dataset also demonstrated consistent synthetic and observed results, the Supplementary Materials show that efficiency can be inflated with increased likelihood for false negatives when this data synthesis approach is used with relatively small samples ($N \leq 30$). Contributors of synthetic datasets derived from relatively small sample sizes should examine measures of efficiency and bias before sharing their synthetic data. This information may also be informative in the context of power analyses and the likelihood that the observed results are replicable.

An additional caution is that subject-level simulant images should not be smoothed, at least for synthetic data that were generated with the imputation approach used in the current study. Each voxel was estimated independently of the adjacent voxels. Spatial smoothing of a synthetic case effectively cancels the variance across adjacent voxels and can produce inflated group-statistics. However, smoothing of the t-score maps can increase the correspondence between observed and synthetic statistical maps, at least for results that have relatively small to medium effect sizes. Smoothing the synthetic statistical maps does not appear to be necessary for large effect sizes and may increase error by extending cluster boundaries and reducing large effects near the peak of a cluster. The code for generating synthetic data includes a function to test smoothing kernel sizes to maximize the correspondence between observed and synthetic results, as implemented in this study. This smoothing kernel information should be conveyed to data recipients because t-score images will require smoothing, rather than the synthetic data.

4.4. Future directions and application

The fully synthetic data approach described here may be useful for researchers who need to analyze restricted-access datasets. For example, researchers could access a secure cloud-based infrastructure to define an imputation model that is aligned with their planned analyses so that the synthetic dataset generated can be analyzed appropriately. This type of approach would likely encourage the contribution of data to a repository by providing a mechanism for safe participant or patient data sharing, and could provide a data resource with greater statistical power than would be achieved by any individual researcher.

A cloud-based data repository composed of data from different research sites or within an imaging center would likely include missingness, which is also common in neuroimaging studies (e.g., susceptibility artifact; Vaden et al., 2012). Future studies could evaluate the current imputation-based approach when there is actual missing data. Multiple imputation can be used to fill-in missing values and accurately recover statistical estimates, including for group-level fMRI analyses (Vaden et al., 2012), when the imputation model contains the same predictors as the analysis model (Loong and Rubin, 2017).

Another area for study is the degree to which other forms of data can enhance the current imputation-based synthesis approach. For example, we considered using Independent Component Analysis (ICA) to generate component maps of the gray matter images and subject-level weights to inform estimation of the gray matter data. ICA could preserve the multivariate associations within an image dataset (Castro et al., 2015). A method to produce component weights rather than individual voxel data could reduce the variability across voxels in synthetic image data. To the extent that component-based data generation accurately represents spatial dependencies across synthetic voxels, smoothing synthetic statistical maps could be unnecessary. Principle Component Analysis (PCA) has also been proposed for generating synthetic datasets to prevent disclosure, in the context of continuous data (Calviño, 2017). Methods such as ICA or PCA may extend the current approach for generating fully synthetic neuroimaging data to multi-modal datasets, as well as enhance the representation of observed univariate and multivariate associations.

Finally, it is also possible that other replacement methods could enhance the quality and usability of synthetic data, while limiting privacy disclosure. For example, reality-based synthetic fMRI data have been used to facilitate power analyses and experiment design (Ellis et al., 2020). Deep learning models have also been developed to create synthetic brain images that appear realistic (Bermudez et al., 2018; Calimeri et al., 2017) to augment training datasets for machine learning methods to perform diagnostic tasks more accurately (Shorten and Khoshgoftaar, 2019). The imputation based approach described here might also be developed to produce single subject data that appear biologically plausible based on simulated informative predictors, although substantial changes would be required to introduce spatial dependence in the generation of image data. There are, however, different goals for deep learning methods like generative adversarial networks and the proposed imputation-based method. The former is focused on creating biologically plausible data while the latter is focused on appropriate statistical inference from the synthetic data.

5. Conclusions

Results from the current study demonstrated that multiple imputation can be used to generate fully synthetic data that accurately replicate results from observed data. This method can be applied to generate synthetic data for other types of experimental and neuroimaging datasets. It may be a useful approach for sharing data when the risk of re-identification and/or potential harm requires a more cautious approach to data sharing compared to de-identifying observed data. This could include data from electronic health records where the code used in this study could be used to generate synthetic datasets based on queried sample parameters and analysis models. Researchers should consider the relative risk of re-identification when sharing data and the approach described here is one way of sharing data when the potential harm from re-identification may be high. Use of the approach described here should be pursued with consideration of its limitations, in particular how sample size, the extent of data replaced per iteration, number of datasets generated, and model discrepancies can affect synthetic results. If the recipient and contributor of synthetic data understand these issues, multiply imputed synthetic data has the potential to enhance scientific integrity, discovery, and education when the observed data cannot be shared.

CRediT authorship contribution statement

Kenneth I. Vaden Jr.: Conceptualization, Methodology, Investigation, Software, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Mulugeta Gebregziabher:** Conceptualization, Writing - review & editing. **Dyslexia Data Consortium:** Resources. **Mark A. Eckert:** Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

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 the Dyslexia Data Consortium and contributors who provided the data for this study.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2020.117284.

References

Abayomi, K., Gelman, A., Levy, M., 2008. Diagnostics for multivariate imputations. J. R. Stat. Soc. Series C: Appl. Stat. 57, 273–291. 10.1111/j.1467-9876.2007.00613.x.

Abramian, D., Eklund, A., 2019. Refacing: Reconstructing anonymized facial features using GANS. IEEE 1104–1108. 10.1109/isbi.2019.8759515.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38, 95–113. 10.1016/j.neuroimage.2007.07.007.

Barnes, S.A., Lindborg, S.R., Seaman, J.W., 2006. Multiple imputation techniques in small sample clinical trials. Stat. Med. 25, 233–245. 10.1002/sim.2231.

Bellovin, S.M., Dutta, P.K., Reitinger, N., 2019. Privacy and synthetic datasets. Stan. Tech. L Rev. 1, 1–51. 10.2139/ssrn.3255766.

Bermudez, C., Plassard, A.J., Davis, T.L., Newton, A.T., Resnick, S.M., Landman, B.A., 2018. Learning implicit brain MRI manifolds with deep learning. Proc. SPIE Int. Soc. Opt. Eng. 10574. 10.1016/j.physbeh.2017.03.040.

Bledsoe, M.J., Russell-Einhorn, M.K., Grizzle, W.E., 2018. Shifting sands: The complexities and uncertainties of the evolving US regulatory, policy, and scientific landscape for biospecimen research. Diagnostic Histopathol. 24, 136–148. 10.1016/j.mpdhp.2017.09.004.

Brakewood, B., Poldrack, R.A., 2013. The ethics of secondary data analysis: considering the application of Belmont principles to the sharing of neuroimaging data. NeuroImage 82, 671–676. 10.1016/j.neuroimage.2013.02.040.

Calimeri, F., Marzullo, A., Stamile, C., Terracina, G., 2017. Biomedical data augmentation using generative adversarial neural networks. In: Lintas, A., Verschure, P.F.M.J., Rovetta, S., Villa, A.E.P. (Eds.), Proceedings of the International Conference on Artificial Neural Networks. Springer International Publishing AG, Cham, Switzerland, pp. 626–634.

Calviño, A., 2017. A simple method for limiting disclosure in continuous microdata based on principal component analysis. J. Off. Stat. 33, 15–41. 10.1515/JOS-2017-0002.

Castro, E., Ulloa, A., Plis, S.M., Turner, J.A., Calhoun, V.D., 2015. Generation of synthetic structural magnetic resonance images for deep learning pre-training. In: Proceedings of the IEEE 12th International Symposium on Biomedical Imaging, Brooklyn. IEEE Computer Society, pp. 1057–1060.

Cocosco, C.A., Kollokian, V., Kwan, R.K.S., Evans, A.C., 1997. BrainWeb: Online Interface to a 3D MRI simulated brain database. NeuroImage 5, S425 10.1016/S1053-8119(97)80018-3.

Dankar, F.K., El Emam, K., Neisa, A., Roffey, T., 2012. Estimating the re-identification risk of clinical data sets. BMC Med. Informat. Decis. Making 12, 1–15. 10.1186/1472-6947-12-66.

Eckert, M.A., Berninger, V.W., Vaden, K.I., 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. 10.1523/ENEURO.0103-15.2015.

Eckert, M.A., Vaden, K.I., Maxwell, A.B., Cute, S.L., Gebregziabher, M., Berninger, V.W.Dyslexia Data Consortium, 2017. Common brain structure findings across children with varied reading disability profiles. Sci. Rep. 7. 10.1038/s41598-017-05691-5.

Eklund, A., Knutsson, H., Nichols, T.E., 2019. Cluster failure revisited: Impact of first level design and physiological noise on cluster false positive rates. Hum. Brain Mapping 40, 2017–2032. 10.1002/hbm.24350.

El Emam, K., Buckeridge, D., Tamblyn, R., Neisa, A., Jonker, E., Verma, A., 2011. The re-identification risk of Canadians from longitudinal demographics. BMC Med. Informat. Decis. Making 11, 1–12. 10.1186/1472-6947-11-46.

Ellis, C.T., Baldassano, C., Schapiro, A.C., Cai, M.B., Cohen, J.D., 2020. Facilitating open-science with realistic fMRI simulation: validation and application. PeerJ. 8, e8564. 10.7717/peerj.8564.

Gorgolewski, K.J., Poldrack, R.A., 2016. A practical guide for improving transparency and reproducibility in neuroimaging research. PLoS Biol. 14, 1–13. 10.1371/journal.pbio.1002506.

Greve, D.N., Fischl, B., 2018. False positive rates in surface-based anatomical analysis. NeuroImage 171, 6–14. 10.1016/j.neuroimage.2017.12.072.

Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., Erlich, Y., 2013. Identifying personal genomes by surname inference. Science 339, 321–325. 10.1126/science.1229566.

Hayasaka, S., Phan, K.L., Liberzon, I., Worsley, K.J., Nichols, T.E., 2004. Nonstationary cluster-size inference with random field and permutation methods. NeuroImage 22, 676–687. 10.1016/j.neuroimage.2004.01.041.

He, Q., Roy, S., Jog, A., Pham, D.L., 2015. An example-based brain MRI simulation framework. Medical Imaging 2015: Physics of Medical Imaging p. 94120P 10.1117/12.2075687 .

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V., Stephan, D.A., Nelson, S.F., Craig, D.W., 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 4, e1000167. 10.1371/journal.pgen.1000167.

Hong, Y.W., Yoo, Y., Han, J., Wager, T.D., Woo, C.W., 2019. False-positive neuroimaging: Undisclosed flexibility in testing spatial hypotheses allows presenting anything as a replicated finding. NeuroImage 195, 384–395. 10.1016/j.neuroimage.2019.03.070.

Kuchinsky, S.E., Vaden, K.I., Keren, N.I., Harris, K.C., Ahlstrom, J.B., Dubno, J.R., Eckert, M.A., 2012. Word intelligibility and age predict visual cortex activity during word listening. Cerebral Cortex 22, 1360–1371. 10.1093/cercor/bhr211.

Loong, B., Rubin, D.B., 2017. Multiply-imputed synthetic data: advice to the imputer. J. Off. Stat. 33, 1005–1019. 10.1515/jos-2017-0047.

Lu, K., 2017. Number of imputations needed to stabilize estimated treatment difference in longitudinal data analysis. Stat. Methods Med. Res. 26, 674–690. 10.1177/0962280214554439.

Manjón, J.V., Coupé, P., Martí-Bonmatí, L., Collins, D.L., Robles, M., 2010. Adaptive non-local means denoising of MR images with spatially varying noise levels. J. Magn. Resonance Imaging 31, 192–203. 10.1002/jmri.22003.

Meng, X.L., 1994. Multiple-imputation inferences with uncongenial sources of input. Stat. Sci. 9, 538–573 doi:10.1214/ss/1177010269.

Nichols, T.E., Das, S., Eickhoff, S.B., Evans, A.C., Glatard, T., Hanke, M., Kriegeskorte, N., Milham, M.P., Poldrack, R.A., Poline, J., Proal, E., Thirion, B., Essen, D.C.Van, White, T., Yeo, B.T.T., 2017. Best practices in data analysis and sharing in neuroimaging using MRI. Nat. Neurosci. 20, 299–303. 10.1038/nn.4500.

Nowok, B., Raab, G.M., Dibben, C., 2016. synthpop: Bespoke creation of synthetic data in R. J. Stat. Softw. 74. 10.18637/jss.v074.i11.

Poline, J.B., Breeze, J.L., Ghosh, S., Gorgolewski, K.F., Halchenko, Y.O., Hanke, M., Helmer, K.G., Marcus, D.S., Poldrack, R.A., Schwartz, Y., Ashburner, J., Kennedy, D.N., 2012. Data sharing in neuroimaging research. Front. Neuroinformat. 6, 1–13. 10.3389/fninf.2012.00009.

Raghunathan, T., Reiter, J., Rubin, D., 2003. Multiple imputation for statistical disclosure limitation. J. Off. Stat. 19, 1–16.

Reiter, J.P., 2005. Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. J. R. Stat. Soc. Series A: Stat. Soc. 168, 185–205. 10.1111/j.1467-985X.2004.00343.x.

Rubin, D.B., 1996. Multiple imputation after 18+ years. J. Am. Stat. Assoc. 91, 473–489. 10.1080/01621459.1996.10476908.

Rubin, D.B., 1993. Statistical disclosure limitation. J. Off. Stat. 9, 461–468.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York, NY 10.2307/3172772.

Scarpazza, C., Tognin, S., Frisciata, S., Sartori, G., Mechelli, A., 2015. False positive rates in voxel-based morphometry studies of the human brain: should we be worried? Neurosci. Biobehav. Rev. 52, 49–55. 10.1016/j.neubiorev.2015.02.008.

Schafer, J.L., 1999. Multiple imputation: a primer. Stat. Methods Med. Res. 8, 3–15. 10.1177/096228029900800102.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6. 10.1186/s40537-019-0197-0.

Song, X., Wang, J., Wang, A., Meng, Q., Prescott, C., Tsu, L., Eckert, M.A., 2015. DeID - A data sharing tool for neuroimaging studies. Front. Neurosci. 9, 1–16. 10.3389/fnins.2015.00325.

Stuart, E.A., Azur, M., Frangakis, C., Leaf, P., 2009. Multiple imputation with large data sets: a case study of the children's mental health initiative. Am. J. Epidemiol. 169, 1133–1139. 10.1093/aje/kwp026.

Tilling, K., Williamson, E.J., Spratt, M., Sterne, J.A.C., Carpenter, J.R., 2016. Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. J. Clin. Epidemiol. 80, 107–115. 10.1016/j.jclinepi.2016.07.004.

Vaden, K.I., Gebregziabher, M., Eckert, M.A., Dyslexia Data Consortium, 2020. Data for: Fully synthetic neuroimaging data for replication and exploration. 10.17632/jtts2d7dtg.1

Vaden, K.I., Gebregziabher, M., Kuchinsky, S.E., Eckert, M.A., 2012. Multiple imputation of missing fMRI data in whole brain analysis. NeuroImage 60, 1843–1855. 10.1016/j.neuroimage.2012.01.123.

Van Buuren, S., Groothuis-Oudshoorn, K., 2011. MICE: Multivariate imputation by chained equations in R. J. Stat. Softw. 45, 1–67. 10.18637/jss.v045.i03.

Wager, T.D., Lindquist, M.A., Nichols, T.E., Kober, H., Van Snellenberg, J.X., 2009. Evaluating the consistency and specificity of neuroimaging data using meta-analysis. NeuroImage 45, S210–S221. 10.1016/j.neuroimage.2008.10.061.

Wechsler, D., 2004. The Wechsler Intelligence Scale for Children (WASI-IV).

Wechsler, D., 1999. Wechsler Abbreviated Scale of Intelligence (WASI).

White, T., Blok, E., Calhoun, V.D., 2020. Data sharing and privacy issues in neuroimaging research: opportunities, obstacles, challenges, and monsters under the bed. Hum. Brain Map. 1–14. 10.1002/hbm.25120.

Yang, J., Fan, J., Ai, D., Zhou, S., Tang, S., Wang, Y., 2015. Brain MR image denoising for Rician noise using pre-smooth non-local means filter. BioMed. Eng. Online 14, 1–20. 10.1186/1475-925X-14-2.