medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A Generative Model For Evaluating Missing Data Methods in Large Epidemiological Cohorts

Lav Radosavljević[1], Stephen M. Smith[2], and Thomas E. Nichols[1]

[1]Nuffield Department of Population Health, University of Oxford
[2]Nuffield Department of Clinical Neurosciences, University of Oxford

April 23, 2024

Abstract

The potential value of large scale datasets is constrained by the ubiquitous problem of missing data, arising in either a structured or unstructured fashion. While there is considerable work on imputation methods, much is focused on small-scale datasets with just tens of variables. When imputation methods are proposed for large scale data, one limitation is the simplicity of existing evaluation methods. Specifically, most evaluations create synthetic data with only a simple, unstructured missing data mechanism, and do not resemble the missing data patterns found in real data. For example, in UK Biobank missing data tends to appear in blocks, because non-participation in one of the sub-studies leads to missingness for all of the sub-study variables.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

1 Introduction

Missing data is common in epidemiological and health data research and presents formidable challenges for many analytical approaches. The causes of missing data vary from being inherent to the study design, to elective non-participation, to faults in measurement etc. Much work has therefore been devoted to evaluating the performance of methods for handling missing data. The most common approaches to comparing imputation methods include simulating data and inducing missingness using a-priori chosen mechanisms[1,2]. Alternatively, artificial missingness is induced on real, complete data[3–5] or real missingness patterns are imposed on simulated data[6]. Simulation studies of this kind usually rely on quite restrictive assumptions that might not be reflective of large scale epidemiological cohorts such as UK Biobank (UKB). For example, while some studies induce missingness in an unstructured manner[1,4], in UKB missing data caused by non-participation in a sub-study/questionnaire comes in "blocks". Specifically, if a subset of participants do not participate in an extension of the core study, then all of these subjects will have missing entries for the variables of this extension, and (when the rows and columns of the subjects-by-variables matrix are suitably reordered) this will form a solid block of missing data. Since UKB and similar datasets consist of many different sub-studies and questionnaires, this is a crucial feature to consider when evaluating the performance of imputation methods.

While there has been work done on the evaluation of existing methods on data with structured missingness and the development of new methods for handling such data, it has been common to use a non data-driven method for inducing structured missingness[2,7,8] or use real missingness patterns imposed on simulated data[6]. Assuming that the structured missingness is Missing Completely at Random (MCAR) is especially problematic in the case where it is created by non-participation, since we know that participants often are disproportionately healthy[9], meaning that the data is not likely to be MCAR. Our aim is therefore to define a method of generating synthetic data which has the same properties as a given medical data set. We want the pattern to satisfy the following three criteria:

1. There is structured missingness[10], blocks of missingness caused by non-participation in sub-studies, as well as unstructured missingness that is not in blocks and is attributable to any other cause.

2. There is an association between inter-variable correlation and inter-variable missingness similarity, typically where tightly correlated variables are more likely to be jointly missing.

3. Missingness is informative in the sense of MAR (Missing at Random), where there is a relationship between missingness in a given variable and the observed elements of other variables.

With our framework for simulating such synthetic data, we evaluate the performance of several imputation methods. We are motivated by associations

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

between the brain imaging variables with health, demographic, behavioural and lifestyle variables in UK Biobank. Thus we consider the subset of $\approx 40\,000$ subjects with imaging derived phenotype (IDP) data, and a collection of $\approx 20\,000$ non-Imaging Derived Phenotypes (nIDPs) variables; these nIDP variables are a mixture of continuous and binary variables (some of the binary variables are 1-hot encoding of categorical variables).

2 Methods

2.1 Terminology

Let n and d be the number of subjects and variables respectively, $\mathbf{X}$ be our $n \times d$ dataset and $\mathbf{M}$ be the $n \times d$ missingness matrix where $M_{ij} = 1$ if variable j is missing for subject i and $M_{ij} = 0$ if it is not missing. The following definitions and notation are central to our work:

1. **Variable-wise missingness pattern**. For any variable $j = 1, 2, ..., d$, the *variable-wise missingness pattern* for variable j is $\mathbf{m}_j^{\mathrm{v}} = (M_{1j}, M_{2j}, ..., M_{nj}) \in \{0, 1\}^n$.

2. **Subject-wise missingness pattern**. For any subject $i = 1, 2, ..., n$, the *subject-wise missingness pattern* for subject i is $\mathbf{m}_i^{\mathrm{s}} = (M_{i1}, M_{i2}, ..., M_{id}) \in \{0, 1\}^d$.

3. **Variable-wise missingness distance.** For any two variables j and j', the *variable-wise missingness distance* between them is the proportion of discordant missingness indicators

$$D_{jj'}^{\mathrm{v}} = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1} \left\{ M_{kj} \neq M_{kj'} \right\},$$

   where $\mathbf{D}^{\mathrm{v}} = D_{jj'}^{v}$ is the $d \times d$ variable-wise missingness distance matrix.

4. **Subject-wise missingness distance.** For any two subjects i and i' the *subject-wise missingness distance* between them is likewise

$$D_{ii'}^{\mathrm{s}} = \frac{1}{d} \sum_{k=1}^{d} \mathbb{1} \left\{ M_{ik} \neq M_{i'k} \right\},$$

   where $\mathbf{D}^{\mathrm{s}}$ is the $n \times n$ subject-wise missingness distance matrix.

5. **Structured missingness**. We call missingness that is caused by non-participation in a sub-study/questionnaire *structured missingness*, resulting in a subset of subjects having missing data for a set of variables. This is also sometimes called block-wise missingness, as when subjects and variables are suitably reordered, this will result in solid blocks of missing data in the data matrix.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

6. **Unstructured missingness**. We call missingness that is not caused by non-participation in a study/questionnaire *unstructured missingness*. This type of missingness will not induce any sort of blocks of missingness.

We now define the stochastic mechanisms that can give rise to missing data. Let $\mathbf{x}$ be a d-dimensional random vector drawn from the same distribution as the data in our data set and $\mathbf{m}$ be its corresponding subject-wise missingness pattern. Let further $\mathbf{x}_{\text{obs}(\mathbf{m})}$ and $\mathbf{x}_{\text{miss}(\mathbf{m})}$ be the observed and unobserved parts of the random vector $\mathbf{x}$ respectively. We follow the terms in Rubin (1976) for different types of missingness:

- *Missing Completely at Random* (MCAR)

$$\mathbb{P}(\mathbf{m}|\mathbf{x}) = \mathbb{P}(\mathbf{m}).$$

  This means that the missingness mask is completely independent from underlying data.

- *Missing at Random* (MAR)

$$\mathbb{P}(\mathbf{m}|\mathbf{x}) = \mathbb{P}(\mathbf{m}|\mathbf{x}_{\text{obs}(\mathbf{m})}).$$

  This means that there exists some dependence between the missingness mask and the underlying data, but that this relationship can be described using only observed data, i.e., the relationship between $\mathbf{m}$ and $\mathbf{x}$ is determined exclusively by the observed part $\mathbf{x}_{\text{obs}(\mathbf{m})}$. For example, this can mean that there exists a group of variables with no missingness which determine the missingness mask $\mathbf{m}$.

- *Missing Not at Random* (MNAR)

$$\mathbb{P}(\mathbf{m}|\mathbf{x}) \neq \mathbb{P}(\mathbf{m}|\mathbf{x}_{\text{obs}(\mathbf{m})}).$$

  For this type of missingness, the relationship between the data and the missingness requires knowledge of underlying data. This is the most difficult setting to handle, assuming no prior knowledge of the mechanism by which missingness is induced, since it has been shown that for any MNAR model explaining missing data in a given data set, there exists an MAR model with equal evidence[11]. In other words, there can be no theoretical guarantees of correctness for MNAR models explaining missing data barring direct knowledge of the missingness mechanism.

Characterising types of missingness is crucial to our work since many methods of handling missing data, most notably Multivariate Imputation by Chained Equations (MICE), have theoretical guarantees under MCAR and MAR[12], while MNAR requires additional assumptions[13].

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2.2 Parameters of the Generative Model

We assume that our data consists of C different sub-studies, where study $c = 0$ is assumed to be a baseline study with no missingness while studies $c = 1, ..., C-1$ are follow up substudies with both structured and unstructured missingness. The following parameters define our generative model:

- $\{\rho\}_{c,c'}$, the distribution of between-variable correlations for all pairs of clusters c, c'. We assume a mixed data generative model[14], where data arise from a multivariate normal distribution with zero means and unit variances: for continuous variables these values are directly observed, while for binary variables the normal variate is latent and the data is obtained by thresholding continuous variables to 0/1. Therefore, $\{\rho\}_{c,c'}$ represents the correlation distribution of the underlying data prior to thresholding.
- π_c the rate of structured missingness for cluster c.
- (α_c, β_c), parameters governing the rate of unstructured missingness for each variable. We assume that the rate of unstructured missingness p_j^{us} is drawn from $\text{Beta}(\alpha_c, \beta_c)$ if feature j belongs to sub-study c.
- Σ_{core}. We assume that d_{core} variables from the baseline study $c = 0$ determine all structured missingness through a logistic model. Σ_{core} is the correlation matrix of these core variables. The core variables are assumed to all be continuous.
- AUC_c, the Area Under the Curve (AUC) score of the logistic model determining structured missingness for sub-study c.

2.3 Estimating Parameters

We estimate the parameters of the model using the following procedure, also detailed in the flowchart in Figure 1.

1. Our C sub-studies are identified using hierarchical agglomerative complete linkage clustering[15].
2. The densities $\{\rho\}_{c,c'}$ are estimated using a histogram for each pair of clusters c, c'.
3. We define a subject i to be structurally missing for a sub-study c if at least 90% of the variables from c are missing for subject i. This will give us the vectors $\mathbf{b}_c^{\text{s}} \in \{0,1\}^n$ where $\mathbf{b}_{c,i}^{\text{s}} = 1$ if subject i is structurally missing for sub-study c. This result also directly gives us π_c, i.e., the probability of a subject having structured missingness for cluster c.
4. Having identified all structured missingness, we can estimate (α_c, β_c) using the method of moments on the remaining, unstructured missingness.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

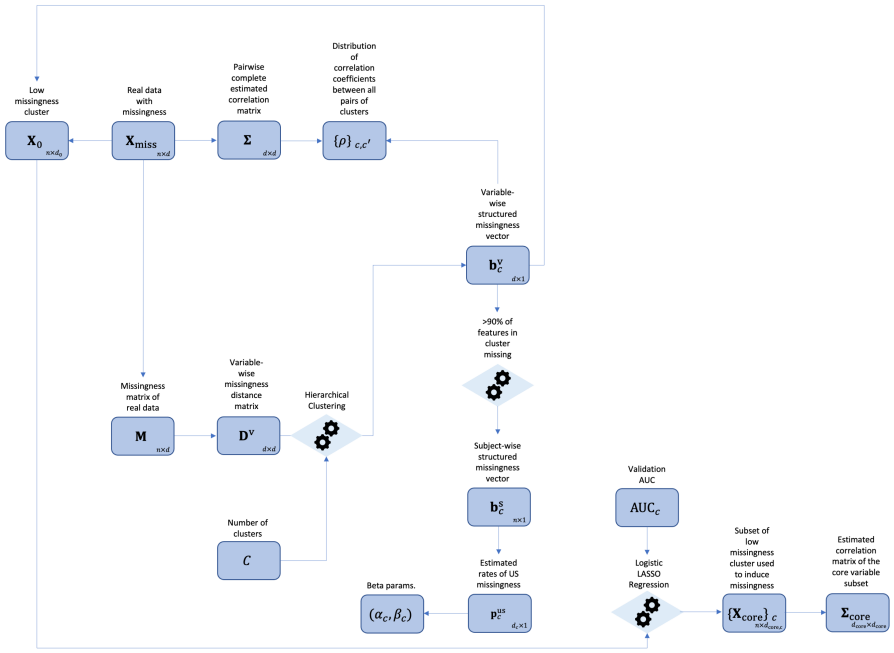

Figure 1: Flow chart of the data analysis pipeline.

5. We use LASSO Logistic Regression (LASSO-LR) [16] to simultaneously identify the core variables that determine structured missingness and AUC_c, by fitting $C-1$ penalised logistic regression models that use the baseline study data $\mathbf{X}_0$ as predictors and the subject-wise structured missingness vectors $\mathbf{b}_c^{\text{s}}$ as outcomes. Specifically, AUC_c is estimated using 5-fold cross validation. Note that the core variables are cluster specific and may or may not overlap for different clusters. We then estimate Σ_{core}, the correlation matrix of the core variables for all substudies. The penalty term λ_c for each LASSO models can be chosen in multiple appropriate ways (see subsection 2.4.1).

2.4 Generating Synthetic Data

The data is generated using a step-wise procedure as seen in Figure 2. Since all continuous variables have unit variance correlation matrices and covariance matrices of continuous data are the same.

1. Using Σ_{core}, we simulate the cluster specific core variables $\{\mathbf{X}^*_{\text{core}}\}_c$ for clusters $c = 1, ..., C-1$, by drawing the full core variable data matrix $\mathbf{X}^*_{\text{core}}$ from $\mathcal{N}(\mathbf{0}, \Sigma_{\text{core}})$.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

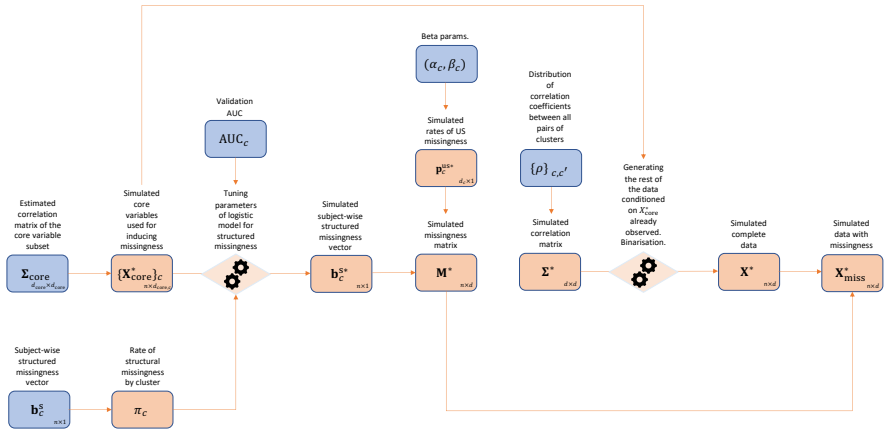

Figure 2: Flow chart of the synthetic data generation pipeline.

2. Using a binary search procedure, we determine intercepts and coefficients of $C-1$ logistic models determining structured missingness, with $\{\mathbf{X}^*_{\mathrm{core}}\}_c$ as covariates, such that the model's AUC score and rate of positive cases will match AUC_c and π_c. All the coefficients of the logistic models are assumed to be equal. Using these models, we generate synthetic subject-wise structured missingness vectors $\mathbf{b}_c^{\mathrm{s}*}$.

3. We generate the rates of unstructured missingness $p_j^{\mathrm{us}*}$ by drawing them independently from $\mathrm{Beta}(\alpha_c, \beta_c)$ where variable j is in sub-study c. Unstructured missingness is assumed to be MCAR and is induced for each subject i and variable j with probability $p_j^{\mathrm{us}*}$. By combining the generated structured and unstructured missingness, we obtain an $n \times d$ synthetic missingness indicator matrix $\mathbf{M}^*$.

4. We simulate the full $d \times d$ correlation matrix Σ^* by drawing its entries from $\{\rho\}_{c,c'}$ and, if necessary, projecting it to the nearest positive definite correlation matrix using Higham's algorithm[17].

5. Having the complete correlation matrix, we generate the rest of the data $\mathbf{X}^*$, conditioned on the already simulated core variable data $\mathbf{X}^*_{\mathrm{core}}$. To allow for binary variables, we threshold a subset of variables to become binary, corresponding to the same number of binary variables in each cluster.

6. Finally the synthetic missingness mask $\mathbf{M}^*$ is imposed upon $\mathbf{X}^*$ to obtain the corresponding synthetic data set with missingness $\mathbf{X}^*_{\mathrm{miss}}$.

This procedure generates synthetic datasets which satisfy the key criteria outlined in the introduction. Crucially, we have access to the true mean vector and covariance matrix, as well as the underlying data obscured by missingness.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2.4.1 Calibrating the Predictability of Missingness

The choice of penalty term λ_c for each LASSO-LR model that predicts structured missingness can most easily be made by selecting the value of λ_c which minimises validation loss. This is, however, not always the option which is most faithful to the assumptions of our generative model. Since our generative model assumes that all core variables have equal importance in predicting structured missingness, we want to choose a value of λ_c which will minimise the number of core variables of low predictive importance, while having a validation loss that is close to that of the optimal value. This is an inevitably arbitrary feature of our generative model and we choose to select a reasonable value of λ_c through trial and error.

2.5 Simulation Study

In order to demonstrate the use of our generative model, we will conduct a simulation study on synthetic data mimicking the UK Biobank Brain nIDPs to evaluate the performance of three commonly used imputation methods on this data set. Our study will test the accuracy of imputation as measured by Mean Squared Error (MSE) for continuous variables and Balanced Accuracy (BA) for binary variables over $B = 20$ synthetically generated datasets. Additionally, to illustrate the difficulty of imputing data with structured missingness, we will perform the same simulation study on copies of the synthetic datasets where missingness has been induced in an MCAR and completely unstructured manner. The data in each sub-study is set to be missing using independent Bernoulli variables with probability equal to the total rate of missingness for the sub-study.

The first imputation method is mean imputation, which will serve as our benchmark method. The second is the matrix completion method SoftImpute[18], which assumes that there exists a low rank approximation of the data set. This method has a tuning parameter, the value of the low rank, which we vary as 5%, 15% and 30% of the full matrix rank. Both the mean imputation and SoftImpute methods will be binarised to impute binary variables using 0.5 as the threshold, so imputed values greater than or equal to 0.5 will be transformed to 1, while the rest will be transformed to 0. The last method is called iterative imputation[19] or ICE, Iterative Imputation by Chained Equations[3], which uses the same iterative procedure as MICE, but does not include randomness in the imputed values and only creates a single imputed data set. By testing the accuracy of iterative imputation we are effectively evaluating the accuracy of the "signal" component of the MICE imputation method. Additionally, MICE is impractical to use in this high dimensional setting with respect to memory use and computational time since it requires a large number of multiply imputed data sets. We chose to impute continuous values using Bayesian Ridge Regression[19] and binary values using Logistic Regression with a Ridge penalty. In high dimensional setting, iterative imputation requires us to choose a subset of $k << d$ variables that will be used to impute each variable j. These variables

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

are normally set to be the k variables with highest absolute correlation with j[19–21], or select the k variables with the most favorable missingness patterns[20]; all else being equal, we favour variables for imputing j which are observed the most often in rows where j is missing and therefore select variables using the rows of the matrix

$$\mathbf{V} = \mathbf{M}^T(\mathbf{1}_{n\times d} - \mathbf{M}), \tag{1}$$

where $V_{jj'}$ is the number of times j' is observed when j is missing.
We propose a third selection method which utilises correlation and missingness jointly, while being applicable to mixed data. It calculates a score $S_{jj'}$ which is proportional to the maximum expected reduction imputation error (MSE for continuous and misclassification rate for binary variables) under the assumption of MCAR and under the generative model described in[14] for joint continuous and binary data, i.e., an underlying multivariate normal distribution with thresholding for binary variables.

- **j and j' are both continuous**

$$S_{jj'} = V_{jj'}\rho^2,$$

  where ρ is the Pearson correlation between variables j and j'.

- **j is continuous and j' is binary**

$$S_{jj'} = V_{jj'}\rho^2,$$

  where ρ is the Pearson correlation between variables j and j'.

- **j is binary and j' is continuous**

$$S_{jj'} = V_{jj'}\left[\int_{-\infty}^{D/\rho_b} \phi(x')\Phi\left(\frac{D-\rho_b x'}{\sqrt{1-\rho_b^2}}\right)dx' + \int_{D/\rho_b}^{\infty} \phi(x')\left(1-\Phi\left(\frac{D-\rho_b x'}{\sqrt{1-\rho_b^2}}\right)\right)dx' - \max\{p, 1-p\}\right]$$

  if $\rho_b > 0$ and

$$S_{jj'} = V_{jj'}\left[\int_{-\infty}^{D/\rho_b} \phi(x')\left(1-\Phi\left(\frac{D-\rho_b x'}{\sqrt{1-\rho_b^2}}\right)\right)dx' + \int_{D/\rho_b}^{\infty} \phi(x')\Phi\left(\frac{D-\rho_b x'}{\sqrt{1-\rho_b^2}}\right)dx' - \max\{p, 1-p\}\right]$$

  if $\rho_b < 0$, where ϕ and Φ are the the probability density function and cumulative distribution function of the standard-normal distribution, ρ is

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

the Pearson correlation between variables j and j', p is the rate of positive cases for variable j, $D = \Phi^{-1}(p)$ and

$$\rho_b = \rho \frac{\sqrt{p(1-p)}}{\phi(D)}.$$

D is the threshold of the underlying standard-normal variable that determines the binary value of j and ρ_b is the correlation between this underlying variable and j'. The reduction in misclassification loss can be calculated directly using these quantities by assuming that we predict 0/1 depending on whether the median of the latent variable conditioned on the value of j' is greater than D or not.

- **j and j' are both binary**

$$S_{jj'} = V_{jj'} \big[p' \max \{ \mathbb{P}(x=1|x'=1), 1 - \mathbb{P}(x=1|x'=1) \} +$$

$$(1-p') \max \{ \mathbb{P}(x=1|x'=0), 1 - \mathbb{P}(x=1|x'=0) \} - \max \{ p, 1-p \} \big],$$

where p and p' are the rates of positive cases for p and p' respectively. Here, the reduction in misclassification loss is calculated directly from the 2×2 contingency table of j and j', since we know the most likely outcome of variable j given the value of j'. This contingency table is calculated using p, p' and ρ.

A formal proof of these results can be found in the supplementary material. We will vary the tuning parameter k to be 10, 50 and 150.

2.6 Illustrative Example: Variable Selection for Predicting Total Grey Matter Volume

In order to demonstrate the validity of the conclusions drawn from our simulation study, we apply our imputation methods to an analytical task on real data and see if there is agreement between the results of the analysis and the conclusions drawn from the simulation study. We chose the task of selecting 15 nIDPs for an Ordinary Least Squares (OLS) model predicting log-transformed normalised total grey matter volume. The total pool to select from is $\approx 15\,000$ nIDPs (nIDPs with 0 variance or missingness above 40% were excluded). Imputation is used here as a pre-processing step and LASSO-LR is used for variable selection. The outcome, i.e., the 15 variables that are selected will vary depending on the imputation method. We compare four different approaches: using only complete varibles, mean imputation, SoftImpute and iterative imputation. The tuning parameters for SoftImpute and iterative imputation are chosen based on their performance in the simulation study. The four approaches are evaluated by the relevance of the 15 selected variables, as measured by the pooled R^2 estimates of each OLS model. To ensure fair assessment of the R^2 scores irrespective of missingness in the selected variables, we use the mice package in R[13] to create $m = 100$ multiply imputed data sets of all selected variables

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 1: Table of cluster sizes.

Cluster	# of Binary Variables	# of Continuous Variables	Total # of Variables
$c = 0$	13458	1227	14685
$c = 1$	371	6434	6805
$c = 2$	131	2012	2143
$c = 3$	3	235	238

and pool the R^2 scores and their standard error estimates for each OLS model according to "Rubin's rules" [22]. We use the following estimator of this standard error [23]:

$$\mathrm{se}(R^2) = \sqrt{\frac{4R^2(1-R^2)^2(n-p-1)^2}{(n^2-1)(n+3)}},$$

where n is the number of observations and p the number of variables.

We also ensure that the baseline variables age squared, sex and Townsend deprivation index are included in the OLS model as potential confounding variables.

3 Results

3.1 Analysis Pipeline

As seen in subsection 2.3, we need to select a value for the number of substudies/clusters C as a parameter of our analysis pipeline. This choice is jointly driven by the data itself as well as the need to select a small number of clusters/substudies to allow us to clearly illustrate our methodology. Figure 3 shows the dendrogram of the 100 last agglomerations in the hierarchical clustering of nIDPs by variable-wise missingness pattern. It can be seen from this dendrogram that $C = 4$ will give us clusters which have a high between-cluster missingness distance relative to within-cluster distance. The choice of $C = 4$ clusters also gives us reasonably sized clusters for our analysis, as seen in Table 1.

The share of each nIDP type by cluster is shown in Figure 4. Cluster $c = 0$ contains almost exclusively health and medical related nIDPs, cluster $c = 2$ contains mostly lifestyle and environment related variables, cluster $c = 3$ almost exclusively contains cognitive phenotype variables and cluster $c = 1$ contains a mix of the remaining types of variables. This results shows that nIDPs of the same type tend to have similar variable-wise missingness patterns.

Figure 5 plots the histograms of the proportions of variable-wise missing data, i.e., fraction of subjects missing for each variable in a cluster. As we can see, the cluster $c = 0$, the cluster that contains mostly health- and medical related variables, has almost no missingness. This is an expected result, as health records usually either contain too much missingness to be included in

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

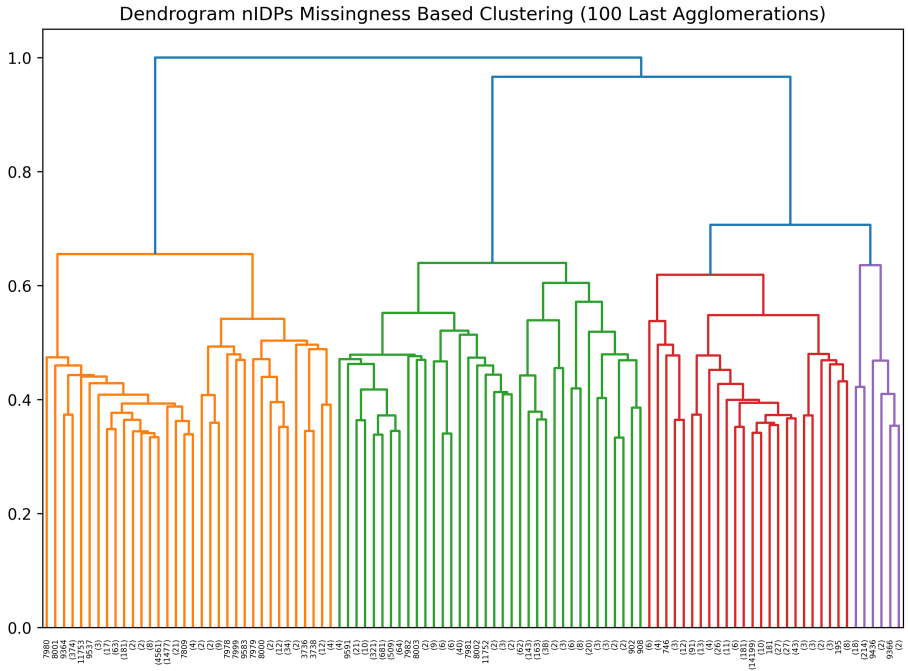

Figure 3: Dendrogram of the 100 last agglomerations in the hierarchical clustering of nIDPs by variable-wise missingness pattern, where distance between merged clusters (y-axis) is the maximal variable-wise missingness distance between agglomerated clusters. We determine that $C = 4$ clusters/substudies is an appropriate choice for illustrating the workings of our method since it gives us reasonably sized clusters with a high between-cluster distance relative to within-cluster distance.

the first place or they have very little missingness as absence of data indicates abscence of recorded disease or diagnosis. Clusters $c = 2, 3$ have intermediate rates of missingness with cluster $c = 3$ having lower rates of missingness as well as a lower variability of rates of missingness, while $c = 1$ has very high rates of missingness. For this reason, we will exclude cluster $c = 1$ from our generative model, as its variables have too high rates of missingness to be interesting to use for imputation.

Figure 6 displays the subject-wise missingness histograms, the proportion of cluster-c variables missing for a given subject. The red line in each plot signifies the 90% threshold for structured missingness, meaning that subjects for which 90% or more of the features assigned to cluster c are missing are considered to have structured missingness for the variables in cluster c. We can see that cluster $c = 3$ has a much clearer separation between structured and unstructured missingness, whereas it is less clear for clusters $c = 1, 2$, likely due to higher rates of unstructured missingness as well as our approximation of $C = 4$ leading to

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

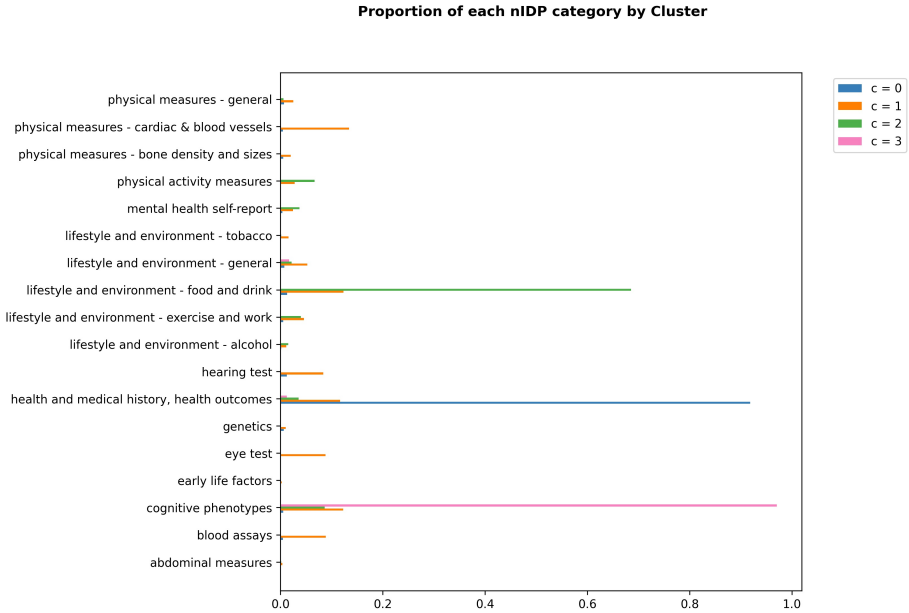

Figure 4: Bar plots detailing the proportion of nIDP variable types in each cluster. Cluster $c = 0$ contains almost exclusively health and medical related nIDPs, cluster $c = 2$ contains mostly lifestyle and environment related variables, cluster $c = 3$ almost exclusively contains cognitive phenotype variables and cluster $c = 1$ contains a mix of the remaining types of variables. This shows that nIDPs of the same type tend to have similar variable-wise missingness patterns.

different clusters being grouped together.

As discussed in subsection 2.4.1, the penalty terms λ_c need to be carefully chosen to not violate the assumptions of our generative model. Manual tuning arrived at a value of $\lambda_c = \exp(6)$ for both clusters $c = 2, 3$ which minimised both the total number of core variables as well as the proportion of binary variables, while having validation AUC scores close to those of the optimal values, as shown in Tables 2 and 3.

The final results of the data analysis are summarised in Table 4. These results indicate that clusters $c = 2, 3$ have similar rates of structured missingness, while cluster $c = 2$ has a much higher rate of unstructured missingness, as indicated by the values of α_c and β_c. It is also apparent that the variables in cluster $c = 3$ have a more informative type of structured missingness as we can see by the higher value of AUC_c.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 2: Table detailing the validation AUC for predicting structured missingness using variables from cluster $c = 0$ and number of binary and continuous variables selected using LASSO Logistic Regression (LASSO-LR) for optimal values of λ_c.

Cluster	AUC_c	# of Binary Variables	# of Continuous Variables	$d_{\text{core},c}$
$c = 2$	0.72	48	575	623
$c = 3$	0.90	271	948	1219

Table 3: Table detailing the validation AUC for predicting structured missingness using variables from cluster $c = 0$ and number of binary and continuous variables selected using LASSO logistic regression using $\lambda_c = \exp(6)$ for both clusters $c = 2, 3$.

Cluster	AUC_c	# of Binary Variables	# of Continuous Variables	$d_{\text{core},c}$
$c = 2$	0.71	5	62	67
$c = 3$	0.86	5	78	83

Table 4: Table of results summarising the data analysis step of the method. These results indicate that clusters $c = 2, 3$ have similar rates of structured missingness, while cluster $c = 2$ has a much higher rate of unstructured missingness, as indicated by the values of α_c and β_c. It is also apparent that the variables in cluster $c = 3$ have a more informative type of structured missingness as we can see by the values of AUC_c.

Cluster	AUC_c	$d_{\text{core},c}$	π_c	α_c	β_c
$c = 2$	0.71	67	0.26	6.0	4.6
$c = 3$	0.86	83	0.29	0.44	4.4

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

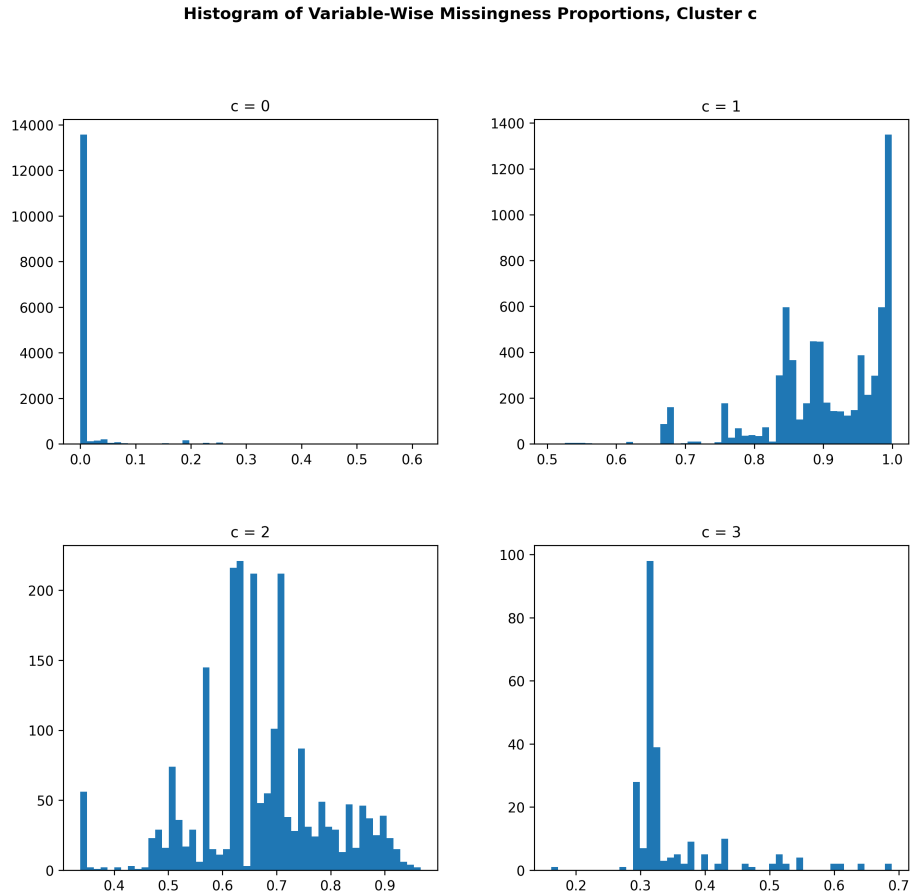

Figure 5: Histograms of the proportions of missing data for variables in each cluster. Each entry in the histogram for cluster c is the proportion of missing data for a single variable belonging to cluster c. Cluster $c = 0$, i.e., the cluster that contains mostly health- and medical related variables, has almost no missingness. Clusters $c = 2, 3$ have intermediate rates of missingness with cluster $c = 3$ having lower rates of missingness as well as a lower variability in the same, while $c = 1$ has very high rates of missingness.

3.2 Simulation Study

Figure 7 plots the imputation accuracy by variable for datasets using the generative model as well as the unstructured equivalent described in sub-section 2.5. The red line in each violin plot is the median of the best performing method in that comparison, i.e., the lowest median MSE and the highest median BA. To interpret the results, it should be noted that the continuous variables of the generative model all have zero mean and unit variance. Iterative imputation that

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

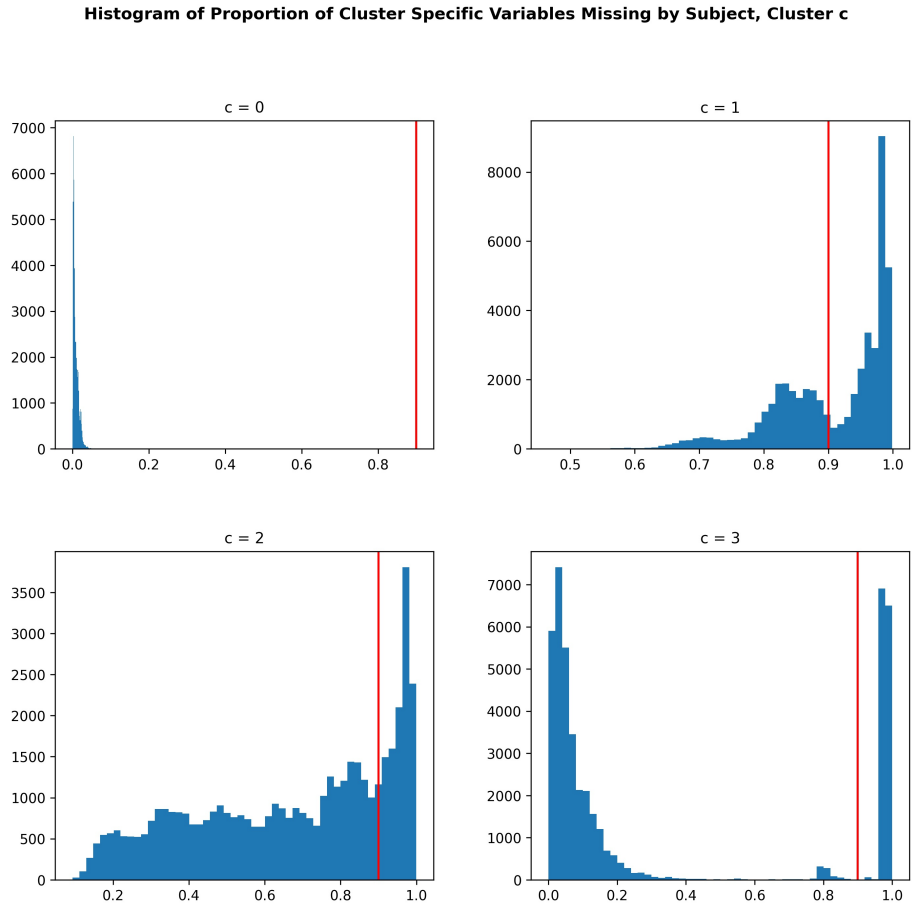

Figure 6: Histograms detailing the proportion of variables assigned to cluster c that are missing, by subject. The red line in each plot signifies the 90% threshold for structured missingness, meaning that subjects for which 90% or more of the features assigned to cluster c are missing are considered to have structured missingness for the variables in cluster c. We can see that the cluster $c = 3$ has a much clearer separation between structured and unstructured missingness, whereas it is less clear for clusters $c = 1, 2$, likely due to higher rates of unstructured missingness as well as our approximation of $C = 4$ leading to different clusters being grouped together.

uses Pearson correlation as its criterion for variable selection is the best performing method overall. It is also notable that the SoftImpute performs poorly for binary data and considerably worse than the iterative imputation methods for continuous data. When comparing the performance between generative model versus completely unstructured missing data, we can see that performance is

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

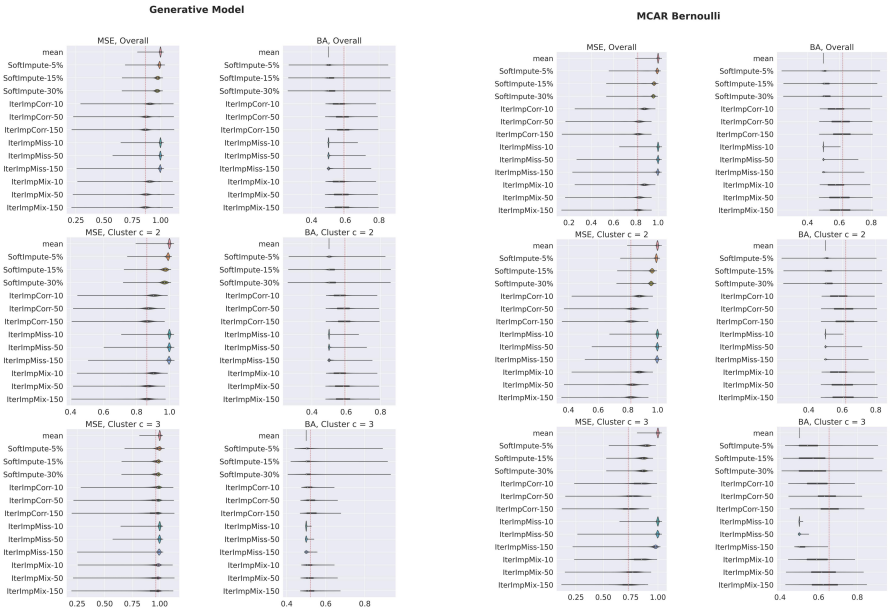

Figure 7: Violin plots of the imputation accuracy by variable. The red line in each violin plot is the median of the best performing method in that comparison, i.e., the lowest median MSE and the highest median BA. Since continuous data is standardised, the MSE scores correspond to 1 minus the variance explained by prediction. It is obvious that iterative imputation that uses Pearson correlation as its criterion for selecting k variables is the best performing method overall. It is also notable that SoftImpute performs poorly for binary data. When comparing the performance between generative model versus completely unstructured missing data, we can see that performance is better for the completely unstructured case, for both continuous and binary variables. This difference is particularly stark for cluster $c = 3$, where there is a lot of structured missingness and very little unstructured missingness, highlighting the difficulty of imputation in this setting.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 5: R^2 scores for the OLS models using the selected variables along with the number of variables in the model that ended up being statistically significant. The results show a modest difference between the three imputation methods, with iterative imputation having the most statistically significant variables and the best R^2 score. The results when using only complete variables are worse with a considerably lower R^2 score as well as fewer variables ending up statistically significant.

Method	R^2 of OLS	se(R^2)	#Statistically Significant in OLS	# Binary vars. selected	Rate of missingness in selected vars.
Complete Variables	0.475	0.0034	11/15	12/15	0.00
Mean	0.510	0.0033	13/15	1/15	0.09
SoftImpute	0.499	0.0034	11/15	1/15	0.11
Iterative Imputation	0.516	0.0033	14/15	2/15	0.10

better for the completely unstructured case, for both continuous and binary variables. This difference is particularly stark for cluster $c = 3$, where there is a lot of structured missingness and very little unstructured missingness, highlighting the difficulty of imputation in this setting. Finally, it should be noted that we are rarely able to explain more than 20% of variance in the missing values and that this could mean that the choice of imputation method will not greatly impact the outcome of many analytical tasks, as the modest accuracy of imputation may not be enough to greatly alter the final outcome.

3.3 Illustrative Example

Table 5 lists the variance explained for the OLS model using the selected variables along with the number of variables in the model that ended up being statistically significant. The results show a modest difference between the three imputation methods, with iterative imputation having the most statistically significant variables and the best R^2 score. The results when using only complete variables are worse with a considerably lower R^2 score as well as fewer variables ending up statistically significant. We also see that the complete variables method selected many more binary variables and this is because the complete variables are mostly health record data, i.e., data assigned to cluster $c = 0$, which is disproportionately binary as seen in Table 1. Meanwhile, the results for iterative imputation are the best, having the highest R^2 score as well as the highest number of statistically significant variables. These results align well with our simulation study; a small difference in the final outcome of the analytical task for different methods caused due to the difficulty of imputing structurally missing data, but with iterative imputation clearly being the best alternative.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

4 Discussion

We have proposed a method for generating large-scale data with complex patterns of missing data that make imputation difficult. In particular, our data-driven simulation framework allows for highly informative missingness and joint missingness for variables that are strongly correlated. This ability to mimic the properties of large scale epidemiological datasets makes our method useful for gaining insight into the performance of handling missing data for different analytical tasks. There are, however, limitations to our model which are important to note and represent potential future work on this topic. Our model assumes multivariate normality for all continuous features, which limits the generalisability of any conclusions drawn from simulation studies of analytical methods that are sensitive to non-gaussianity or strong outliers. In such scenarios, it is possible that conclusions drawn using our generative model would unduly favour linear methods over more complicated black-box methods that would fare better on real, non-Gaussian data. This could be solved by parameterising the model differently, allowing for more flexibility on the underlying multivariate distribution, or by using non-parametric methods.

Another potential limitation of our generative model is that we assume that the correlation structure of the data closely follows the missingness structure. This is because we assume that for any pair of clusters c, c', the correlation between pairs of variables in c and c' are drawn independently from some distribution $\{\rho\}_{c,c'}$, i.e., we assume that there is no further covariance structure within or between sub-studies. We have found this to be approximately true for the nIDPs that we have been working with, but this might not be the case for other datasets. This could be solved by modelling missingness and correlation structure jointly in a way which allows for further complexity inside sub-studies.

In this paper we chose to use $C = 4$ clusters of variables as an approximation of reality in order to be able to inspect the properties of these clusters separately. In all likelihood, the true number of substudies is higher, and more representative results could be obtained by choosing a higher figure. We deemed it necessary to use a lower number in order to demonstrate the inner workings of our method. When allowing C to be higher, we found $C = 15$ clusters with 100 or more variables present in the data set. These clusters all had a very clear separation between the structured and unstructured missingness, which bolsters our hypothesis that missingness in UKB data can be effectively modelled as we have suggested.

The overall conclusion that can be drawn from the results of our simulation study as applied to nIDPs from the UKB Brain Imaging cohort is that imputation using commonly used methods for high dimensional imputation problems is incredibly difficult and might yield similar results as using mean imputation. This was backed up by the results from the illustrative example. Even though the topic of developing such methods is considered to be beyond the scope of this paper, the results do not indicate that the development of such methods is a fruitless exercise, as the methods that were included in the paper are not specifically tailor made to handle structured missingness. This means that fu-

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

ture work on this topic will include the development of suitable methods for handling missing data in this difficult setting.

Supplementary Material

All code can be found in the following GitHub repository: https://github.com/lavrad99/Generative_Model_Missing_Data/

Acknowledgements

LR is supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1)

The Wellcome Centre for Integrative Neuroimaging (WIN FMRIB) is supported by core funding from the Wellcome Trust (203139/Z/16/Z).
SS: Wellcome Trust Collaborative Award 215573/Z/19/Z
The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health

References

[1] Le Morvan M, Josse J, Moreau T, Scornet E, Varoquaux G. NeuMiss networks: differentiable programming for supervised learning with missing values. Advances in Neural Information Processing Systems. 2020;33:5980-90.

[2] van Loon W, Fokkema M, de Rooij M. Imputation of missing values in multi-view data. arXiv preprint arXiv:221014484. 2022.

[3] Jarrett D, Cebere BC, Liu T, Curth A, van der Schaar M. Hyperimpute: Generalized iterative imputation with automatic model selection. In: International Conference on Machine Learning. PMLR; 2022. p. 9916-37.

[4] Ghalebikesabi S, Cornish R, Holmes C, Kelly L. Deep generative missingness pattern-set mixture models. In: International Conference on Artificial Intelligence and Statistics. PMLR; 2021. p. 3727-35.

[5] Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and managing missing structured data in electronic health records: data analysis. JMIR medical informatics. 2018;6(1):e8960.

[6] Gentry AE, Kirkpatrick RM, Peterson RE, Webb BT. Missingness adapted group informed clustered (MAGIC)-LASSO: a novel paradigm for phenotype prediction to improve power for genetic loci discovery. Frontiers in Genetics. 2023;14.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[7] Jackson J, Mitra R, Hagenbuch N, McGough S, Harbron C. A Complete Characterisation of Structured Missingness. arXiv preprint arXiv:230702650. 2023.

[8] Little TD, Jorgensen TD, Lang KM, Moore EWG. On the joys of missing data. Journal of pediatric psychology. 2014;39(2):151-62.

[9] Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. American journal of epidemiology. 2017;186(9):1026-34.

[10] Mitra R, McGough SF, Chakraborti T, Holmes C, Copping R, Hagenbuch N, et al. Learning from data with structured missingness. Nature Machine Intelligence. 2023;5(1):13-23.

[11] Molenberghs G, Beunckens C, Sotto C, Kenward MG. Every missingness not at random model has a missingness at random counterpart with equal fit. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2008;70(2):371-88.

[12] Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? International journal of methods in psychiatric research. 2011;20(1):40-9.

[13] Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2011;45:1-67.

[14] Demirtas H, Doganay B. Simultaneous generation of binary and normal data with specified marginal and association structures. Journal of Biopharmaceutical Statistics. 2012;22(2):223-36.

[15] Sorensen TA. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biol Skar. 1948;5:1-34.

[16] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1996;58(1):267-88.

[17] Higham NJ. Computing the nearest correlation matrix—a problem from finance. IMA journal of Numerical Analysis. 2002;22(3):329-43.

[18] Mazumder R, Hastie T, Tibshirani R. Spectral regularization algorithms for learning large incomplete matrices. The Journal of Machine Learning Research. 2010;11:2287-322.

[19] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.23.24306030; this version posted April 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[20] Van Buuren S, Oudshoorn K. Flexible multivariate imputation by MICE. Leiden: TNO; 1999.

[21] Graham JW. Missing data analysis: Making it work in the real world. Annual review of psychology. 2009;60:549-76.

[22] Rubin DB. Multiple imputation for nonresponse in surveys. Wiley Series in Probability and Statistics. Wiley; 1987.

[23] Olkin I, Finn JD. Correlations redux. Psychological Bulletin. 1995;118(1):155.