RE-CALIBRATED WASSERSTEIN GAN FOR LARGE-SCALE IMPUTATION WITH INFORMATIVE MISSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Missing data are pervasive in electronic health records (EHR) and oftentimes the missingness is informative (i.e. Missing Not At Random). Presently available imputation methods typically do not account for this informative missingness or are computationally infeasible to handle the scale of EHR data. We develop a deep learning imputation method based on *recalibrating* a Wasserstein Generative Adversarial Network (WGAN) to account for informative missingness in high-dimensional quantitative medical data. We propose a new quantile re-weighting technique to ensure distributional equivariance under informative missingness and integrate it with WGAN to enable efficient imputations in large-scale observational data in presence of informative missingness and covariate imbalance. Results from our proposed algorithm show better recovery compared to present methods in both synthetic and real-world data from the Reactions to Acute Hospitalization (REACH) and laboratory test results of COVID-19 patients in the New York Metropolitan area from the INSIGHT dataset.

1 INTRODUCTION

In the wake of the COVID-19 pandemic, the responsible use of big data for healthcare is an especially urgent and important issue. Large-scale *electronic health records* (EHR) have been increasingly used to inform clinical decisions and discover disease etiologies. Although they contain unprecedentedly rich information, EHR are incomplete and sporadic records of individual patients' health conditions. When a patient visits a healthcare facility, variables recorded during that visit often depend on their health conditions. Such missingness in variables are *informative* as they are directly linked with the patient's underlying health conditions. For example, patients with diabetes will have more frequent glucose records compared to non-diabetic patients. A subject's systolic blood pressure or heart rate is less likely to be measured if it is low Yoon et al. (2018b). In psychological questionnaires, variables that measure the severity of depression or mania may be correlated to the likelihood of response. Another key feature of EHR is its high dimensionality; this attribute also conceals multifarious hidden relationships between different variables and their missingness.

There are several ways to make missing data amenable for analysis. One simple way is by deleting entire vectors of variables or subjects (Silva and Zárate (2014)). However, this approach can remove important aspects of the data and lead to compounded errors with unintended consequences (Graham (2009)). Imputation, or filling in missing values with plausible predicted values from a given model that uses the existing observed variables, is another approach. Imputation using deep learning (DL) have been a recent trend due to its computational advantages: many of these method assume that data are *missing completely at random* (MCAR), where missingness is independent from the data, or only dependent on observed data (*missing at random* (MAR)) (Yoon et al. (2018a); Li et al. (2019); Yoon and Sull (2020)). However, in the real world, this assumption is not necessarily correct for analyzing EHR (Albers et al. (2018)), as covariates are often not MAR.

These problems of informative missingness (IM) have been addressed in inverse probability weighted (IPW) approaches from nonparametric statistics (Wei et al. (2012); Yuan and Dong (2019); Xie and Zhang (2017)). The advantage of DL methods over these above methods, however, is that they are faster and can handle "bigger" data, but they also suffer disadvantages in their inaccuracy and inability to extend beyond the MCAR realm. Wang et al. (2021) demonstrate that traditional methods like *multiple imputation by chained equations* (MICE) outperforms DL based approaches for small to moderately sized datasets. However, MICE is much slower. To account for these gaps in imputation methodology, we introduce a novel method based on WGAN with improved accuracy in

recovering the distributive properties of the missing data. Our **novel contributions** are twofold: (1) introduction of a new WGAN-based imputation method whose objective function is recalibrated by estimated missingness probabilities to account for the drawbacks of the Wasserstein distance (noted by Stanczuk et al. (2021); Fedus et al. (2017)), (2) application to high-quality real EHR data with quasi-experimental missingness settings, together with simulations that approximate them.

2 Preliminaries and Related Work

We describe several existing methods that are commonly used for imputing missing data. One commonly used approach is MICE. MICE fits regression models for each variable that has missing values after conditioning on all other observed variables (Breiman et al. (1983); van Buuren and Groothuis-Oudshoorn (2011)). MICE is regarded as a state-of-the-art approach, but the biggest limitation is that computation time increases quadratically as with the number of variables (Wang et al. (2021)). Another class of methods use neural networks (NN). MIDA is one example based on denoising autoencoders (Gondara and Wang (2017)). *Generative adversarial networks* (GANs) are another popular framework for imputation (Goodfellow et al. (2014)); examples include GAIN and WGAIN (Friedjungová et al. (2020); Yoon et al. (2018a); Li et al. (2019)). Luo et al. (2018) proposed a GAN-based imputation in time series that also implicitly account for time-serial informative missingness. In the following section, we describe GAN and its Wasserstein variant in detail.

2.1 (Wasserstein) Generative Adversarial Networks

Broadly speaking, GANs train two neural networks *generator* G and *discriminator* D to compete against each other to reach a global maximum. Talas et al. (2020) describe GAN as an *evolutionary arms race* between two parties wherein one keeps trying to outdo the other. G tries to generate a distribution P_G that is as similar to 'real' distribution P_{data}. D tells if what is generated by G comes from P_{data} rather than P_G. Wasserstein GAN (WGAN) is an alternative GAN that yields demonstrable advantages over the original by minimizing the Wasserstein distance between the P_{data} and P_G (Arjovsky et al. (2017)).

Consider a p-dimensional space $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_p$, let (x_{data}, x_G) be elements in $\mathcal{X}$ whose marginal distributions are P_{data}, P_G. $\gamma(x_{\text{data}}, x_G)$ indicates how much mass must be transported from x_{data} to x_G in order to transform the distribution P_{data} *into* P_G. The Wasserstein-1 distance W_1 is the minimum cost of the *optimal transport* plan between the distributions (details in Appendix A (Villani (2008))). The competing objectives of G and D is motivated by the Kantorovich Duality, which restates the Wasserstein distance as the supremum of differences in expectations of densities with Lipschitz functions f as follows:

$$W_1(P_{\text{data}}, P_G) \equiv \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim P_{\text{data}}}[f(x)] - \mathbb{E}_{x_G \sim P_G}[f(x_G)]. \tag{1}$$

The overall objective of WGAN is to minimize W_1 (we use W_1 throughout the manuscript without loss of generalization for $p > 1$) for P_{data} and P_G. (W)GANs are popular tools of imputation because they are fast, but they also naturally encourage multiple imputation. Yoon et al. (2018a) and Yang et al. (2019) in particular, posit that GAN-based approaches can be considered forms of *multiple imputation* because its generative aspect "models all features with missing values simultaneously". However, they are not without its limitations. Wang et al. (2021) demonstrates that MICE actually outperforms nearly all NN-based approaches for small to moderately sized datasets , but its computing time increases quadratically compares to DL methods (Gondara and Wang (2017)). Stanczuk et al. (2021) remark that existing WGAN methods do not necessarily minimize the Wasserstein distance or KL divergence (also noted by Fedus et al. (2017)). Furthermore, even if W_1 is minimized, Stanczuk et al. (2021) demonstrate that it is not even necessarily a desirable criterion for generative modeling. Li et al. (2017) note that few (W)GANs guarantee theoretical convergence except under very specific conditions. Stanczuk et al. (2021)) attribute the effectiveness of WGAN instead to the D regularizer in the gradient penalty implementation (Gulrajani et al. (2017)).

We propose a method that more accurately imputes the data distribution by learning from the observed data as well as its estimated missingness. We aim to " preserv[e] the distributional characteristics of the data" (Little and Rubin), by using a *recalibrated* Wasserstein-like distance that is reweighted by the estimated missing probability to achieve better fidelity and minimize bias in reconstruction. We describe this in Section 3. Noting the absence of theoretical guarantees for WGAN minimization, we are guided by a practical approach to *empirically* best redistribute observed data and imputations using estimated missingness probabilities. In Section 5 we demonstrate that our

proposed method outperforms other existing methods in (ordinary) W_1 distance minimization on real data, following evaluation of simulations in Section 4.

2.2 NOTATION

Suppose that $\mathbf{X} = (X_1, ...X_p)$ are random variables in $\mathcal{X}$ (defined in Section 2.1 with n observations with missing entries. For each j-th index, let X_j^{obs} represent the section of the variable that is observed, and let X_j^{msg} be the subset of the vector that is missing. The mask matrix $\mathbf{M} = M_1, ..., M_p$ takes values of 1 if the value of X_{ij} is observed, and 0 if it is missing. Each M_j is the indicator *vector* for each column (feature), and M_{ij} represents the specific entry at the i-row and j-th column. Let X_j and $\mathbf{X}$ represent the data vector/matrix as defined in previous sections. Note that $\mathbf{X}$ is the matrix of observations with missing entries filled in with uniform noise $\mathbf{U}$. G is the generator with associated parameters Θ_G and D is the discriminator with Θ_D. A flowchart of the relationships between these variables is found in Figure 1.

3 PROPOSED METHOD

We use the model framework for imputation proposed by Yoon et al. (2018a) (GAIN). The main features of the model architecture are generator G and D. The generator generates *fake data* (i.e. imputations) and the discriminator judges the quality of the generation and updates its ability to discern between the synthetic and the real data. D estimates the probability of mask $\mathbf{M}$ (Yoon et al. (2018a)), and the quality of the estimate of $\mathbf{M}$ is used to score D. D and G each have 3 layers (details in Appendix C.1). All inputs and outputs of G and D are matrices with dimensions $n \times p$.

In each iteration of our proposed method, random noise $\mathbf{U}$ is added into the observed data matrix *with missing* $\mathbf{X}_{NA}$: $\mathbf{X} = \mathbf{M} \odot \mathbf{X}_{NA} + (1 - \mathbf{M}) \odot \mathbf{U}$. Then $\mathbf{X}$ is fed into generator G: $G(\mathbf{X}|\Theta_G) := \dot{\mathbf{X}}$, which denotes the *purely* imputed dataset, of which the observed entries are also imputations. $\dot{\mathbf{X}}$ takes fake synthetic values for the entries which are observed as well. $\hat{\mathbf{X}}$ is matrix of the mixture of observed and imputed missing values:

$$\hat{\mathbf{X}} = \mathbf{M} \odot \mathbf{X} + (1 - \mathbf{M}) \odot \dot{\mathbf{X}}. \tag{2}$$

Figure 1 (the other parts of the diagram is described in the following sections) shows the differences between the three types of data matrices $\mathbf{X}$ (input with random noise), $\dot{\mathbf{X}}$ (pure imputation), and $\hat{\mathbf{X}}$ (composite). These data matrices are fed into D, each with outputs $\tilde{\mathbf{M}}$, $\dot{\mathbf{M}}$, $\hat{\mathbf{M}}$ that take on estimated *probabilities* between 0 and 1 (as in Yoon et al. (2018a)). $\dot{\mathbf{M}} = D(\dot{\mathbf{X}}|\Theta_D)$ represents the "fake" $\mathbf{M}$ estimate of imputations (using the terminology of Arjovsky et al. (2017), though $\mathbf{X}$ is a mixture of observed and random noise), while $\tilde{\mathbf{M}} = D(\mathbf{X}|\Theta_D)$ represents the "real" (i.e. real observations plus noise) estimates. Their difference are assessed in relation to $\mathbf{M}$ to train D. $\hat{\mathbf{M}} := D(\hat{\mathbf{X}}|\Theta_D)$ represents the *composite* (i.e. *best guess*) and will serve as the probability weights for the quantile reshaping adjustment, which will be described in Section 3.1.

3.1 MISSINGNESS ASSUMPTIONS

Prior work has predominantly focused on the the MAR and MCAR cases (as mentioned in Section 2), citing the difficulty of correcting for MNAR. We wish to impute missing values in data that is not at random. We make two assumptions about missingness. The first is one of *informative missingness* (IM) within a variable. An example is a variable that is be more likely to be un-recordable if its value (inclusive of missings) is very high. The second is that of *covariate imbalance* (CI), where the other variables inform the missingness of the variable X_j itself. We adjust empirical cumulative distribution functions (CDF) of *observed* and *missing* data adjusted by their estimated missingness probabilities $\hat{\mathbf{M}}$. We note that the weighted summation of such CDFs are used in prior work in multiple imputation from the nonparametric statistics perspective (Wei et al. (2012)). If variable X_j is IM, then $\mathbb{P}(X_j \text{ is missing}) = \mathbb{P}(M_j = 0|X_j)$. To correct for such dependency, we model the missingness of each variable with the output $\hat{\mathbf{M}}$ of D,

$$\mathbb{P}(X_j \text{ is missing}) := 1 - D(\hat{\mathbf{X}}|\Theta_D)_j := 1 - \hat{M}_j.$$

To impose distributional equivariance between the missing and observed variables (Section 3.2) we reweight empirical CDFs of each variable with missingness probabilities $\hat{M}$ iteratively. More details on these assumptions are found Appendix B.

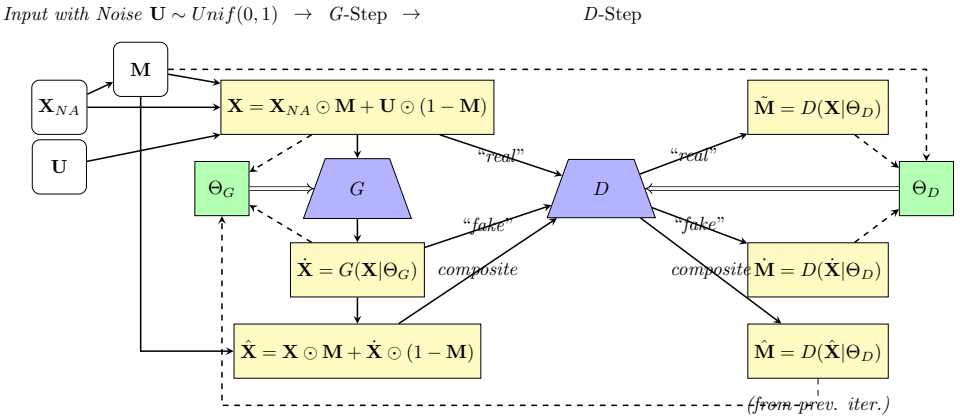

Figure 1: Diagram of a single iteration of the proposed algorithm. The diagram flows in order from left to right, where the same $\mathbf{X}_{NA}$ is combined with a randomly generated $\mathbf{U}$. First the G step is applied to generate $\dot{\mathbf{X}}$and D step on right. Blue trapezoids represent the NNs G and D with associated parameters Θ_G and Θ_D. White boxes represent *raw* data, yellow boxes represent synthethic variables that are used iteratively to train parameters. Green boxes represent the parameters to be trained using stochastic gradient descent (ADAM algorithm) from *Tensorflow*. Dashed arrows represent variables used to train the parameters. Input $\tilde{\mathbf{M}}$ to G from the *prior* step from the D optimization. (" $\Longrightarrow$ ") signifies the fixing parameters Θ_G, Θ_D for NNs G, D.

3.2 Quantile Recalibration Method for Distributional Equivariance

We develop a method to reshape the empirical distributions of data with missing entries. We follow the approach of Yuan and Dong (2019), who reweigh empirical likelihoods by missingness probabilities, and *recalibrate* the Wasserstein distance between observed and missing (imputed) components by reweighing it by estimated missingness probabilities (3). For fixed j, let X_j be the jth variable and let M_j be its associated mask vector. For any of the variables $\mu_{ij} = \pi_{ij}$, M_{ij}, X_{ij}, V_{ij}, n_j,q_k (to be defined within Algorithm 1), let μ^*_{ij} denote either μ^{ob}_{ij} or μ^{msg}_{ij}. This method re-weights each

Algorithm 1 Quantile reshaping method for data vectors with missing entries

Input: Data vector X_j with missing entries

1. Partition jth variable X_j into components $X^{\text{obs}}_j := X_j|_{M_j=1}$ and $\hat{X}^{\text{msg}}_j := \hat{X}_j|_{M_j} = 0$. Associated estimated missingness probabilities are written as $\pi^*_{ij} = \hat{M}^*_{ij} = D(\hat{\mathbf{X}}|\Theta_D)_{ij}\big|_{M_j}$
2. For each j, define each V^*_{ij} (with associated length n^*_j): $V^*_{ij} := 1/\pi^*_{ij} \Big/ \sum_{i=1}^{n^*_j} 1/\pi^*_{ij}$
3. Now we take the observed and imputed (if missing) values of variable X_j. *Rank* each X^*_j (in the observed or missing regime) and then define each $V^*_{(l)j}$ as V^*_{ij} that corresponds with the index of $X^*_{(l)j}$, the l-th ranked X^*_j. Define each quantile *height* q^*_k ($*$ for either observed or missing) to be:

$$q^*_1 = V^*_{j(1)} \Big/ \sum_{l=1}^{n} V^*_{j(l)}, ..., q^*_h = \sum_{k=1}^{h} V^*_{j(k)} \Big/ \sum_{l=1}^{n^*_j} V^*_{j(l)}, ..., q^*_{n^*-1} = \sum_{k=1}^{n^*-1} V^*_{j(k)} \Big/ \sum_{l=1}^{n^*_j} V^*_{j(l)}, 1$$

4. For a range of τ_k, where $k = 1, ..., K$, the quantile functions $\tilde{Q}^*$ (for either observed or missing) can then be calculated as follows :

$$\tilde{Q}^*(\tau_k, X_j|M_j) = \inf\left\{X^*_{(k),j} \text{ s.t. } q^*_k \leq \tau\right\}$$

Choose evenly gridded quantiles $\boldsymbol{\tau} = \{\tau_1, ..., \tau_K\}$, then linearly interpolate the K points.

Calculate $\tilde{Q}^*(\tau_k, X_j|M_j)$ across all $\tau_1, ..., \tau_K$ for observed and missing variables, calculate the norms of their difference, $\|\tilde{Q}(\tau_k, X_j|M_j = 1) - \tilde{Q}(\tau_k, \dot{X}_j|M_j = 0)\|$, then add to obtain (3).

empirical CDF of each variable based on its probability of being observed. This technique is known as *inverse probabilty weighting* (IPW) when it is applied to reweigh coefficients;similar approaches have been used in quantile regression settings (Cheng and Wei (2018); Seaman and White (2013)). We adjust the overall likelihood (W_1 in WGAN) to encourage distributional equivariance between

observations and imputations. For each vector X_j with mask vector M_j; each i-th entry is X_{ij} and M_{ij} respectively. Let $\mathbf{\Phi}_X(x, \boldsymbol{\omega}|M_j)$ be a *generalized conditional* CDF that allows for reweighting of different quantile segments given mask vector M_{ij} where $\boldsymbol{\omega}$ is the weight of each indicator (Wei et al. (2012)). We use the "composite" estimate of $\hat{M}_j = D(\hat{\mathbf{X}}|\Theta_D)$ and $\mathbf{x}_j$ represents an array of values that correspond to a grid of evenly-spaced inverse quantiles for each X_j. This grid approach is used often as a density redistribution method in quantile regression (Wei and Yang (2014)),

$$\mathbf{\Phi}_{X_j}(x, \boldsymbol{\omega}|M_j = 1) = \frac{\sum_{i=1}^{n} \omega_i \mathbf{1}(X_i \leq x)\mathbf{1}(M_{ij} = 1)}{\sum_{i=1}^{n} \omega_i \mathbf{1}(M_{ij} = 1)}.$$

The overall loss of the adjusted Wasserstein Distance $\mathcal{W}$ is

$$\mathcal{W}(P_{\text{data}}, P_G, \Theta_G) = \frac{1}{p}\sum_{j=1}^{p} \left\| \mathbf{\Phi}^{-1}_{X_j}(\mathbf{x}_j, \hat{M}_j|M_j = 1) - \mathbf{\Phi}^{-1}_{G(\mathbf{X},\Theta_G)_j}(\mathbf{x}_j, 1 - \hat{M}_j|M_j = 0) \right\|, \quad (3)$$

We write $\dot{\mathbf{X}}$ as $G(\mathbf{X}, \Theta_G)$ to emphasize the optimization of parameter Θ_G. We posit that $\mathcal{W}(\cdot)$ is more useful than the canonical W_1 for imputation in that it rebalances the missing data using the iterative estimates of the missingness probabilities. We show later in Section 5 that the method does empirically minimize W_1 for real data as well.

3.3 Objective Functions

We introduce an imputation method that uses the method of quantile reshaping to correct for informative missingness which we call Recalibrated Wasserstein Imputation GAN (RWIGAN). The criterion for RWIGAN is comprised of two parts: maximizing the D loss and minimizing the G loss. These steps are alternated and a schematic diagram for the implementation is shown in Figure 1. When training D, the differences between true mask $\mathbf{M}$ and "real" (i.e $\mathbf{M} - \tilde{\mathbf{M}}$) and "fake"($\mathbf{M} - \dot{\mathbf{M}}$) estimates are made as as far apart as possible. The goal of D is to make the *fake* guess as bad as possible (i.e. in competition with G). First, we define a monotone transformation for the output probability estimate of $D(\cdot)$ that better captures its distance from its ground-truth value $\mathbf{M}$,

$$\mathbb{E}[D'(\mathbf{X}, \Theta_D)|\mathbf{M}] = -\sum_{i=1}^{n}\sum_{j=1}^{p} \text{logit}\big((D(\mathbf{X}, \Theta_D)_{ij} - M_{ij})^2\big). \quad (4)$$

We apply the logit transformation to the squared difference between $\mathbf{M}$ and each of $D(\mathbf{X})$ and $D(\dot{\mathbf{X}})$ to obtain a normalized value to better train the D loss which is the difference of the two expectations. These loss functions are written as follows:

$$\max_{\Theta_D} \mathcal{L}_D(\mathbf{X}, \mathbf{M}, \Theta_D) = \max_{\Theta_D, \|\Theta_D\|_L \leq 1} \mathbb{E}_{\mathbf{X}\sim P_{\text{data}}}[D'(\mathbf{X}, \Theta_D)|\mathbf{M}] - \mathbb{E}_{\dot{\mathbf{X}}\sim P_G}[D'(\dot{\mathbf{X}}, \Theta_D)|\mathbf{M}], \quad (5)$$

$$\min_{\Theta_G} \mathcal{L}_G(\mathbf{X}, \mathbf{M}, \Theta_G) = \min_{\Theta_G} \mathcal{W}(P_{\text{data}}, P_G, \Theta_G). \quad (6)$$

The logit transform aligns with the gradient direction of G minimization as per the primal formulation of the W_1 ((8) in Appendix A) before it is transformed into its Kantorovich Dual (Villani (2008)). The D-loss is constructed from the assumption that the distance between $\mathbf{M}$ and $\dot{\mathbf{M}}$ is greater than that between $\mathbf{M}$ and $\tilde{\mathbf{M}}$, and the bigger this difference is, the better D is able to discriminate, as G in turn becomes better at generating samples.

Gulrajani et al. (2017)'s WGAN uses a gradient penalty (WGAN-GP) to enforce the Lipschitz constraint while training D; following this, we augment the $D-$loss with a penalty parameter to $\mathcal{L}^{\lambda}_D(\mathbf{X}, \mathbf{M}, \Theta_D)$ during the training steps. For $\mathbf{Y} = \xi\mathbf{X} + (1 - \xi)\dot{\mathbf{X}}$, and ξ is randomly drawn from a uniform (0,1) distribution,

$$\begin{aligned}\mathcal{L}^{\lambda}_D(\mathbf{X}, \mathbf{M}, \Theta_D) = \mathbb{E}_{X\sim P_{\text{data}}(X)}[D'(\mathbf{X}, \Theta_D)|\mathbf{M}] - \mathbb{E}_{\dot{\mathbf{X}}\sim P_G}[(D'(\dot{\mathbf{X}}, \Theta_D)|\mathbf{M}] \\ + \lambda\mathbb{E}_{\mathbf{Y}}[\|\nabla_{\Theta_D} D'(\mathbf{Y}, \Theta_D)\|_2 - 1)^2|\mathbf{M}].\end{aligned} \quad (7)$$

The G loss is equivalent to minimizing $\mathcal{W}$ (3). We constrain the gradient to be positive to ensure monotonocity for global convergence; we demonstrate that the objective is monotonic and also show empirical convergent properties in Appendix C.3. In the implementation of the algorithm, we set the batch size to be 100 for the simulations (around one-third of n) and set $\lambda = 1$. Differing settings of λ and batch size (the only free parameters of the method) do not alter the outcome in a significant way. We cycle a single iteration of D, though more than one is possible for smoother training (as is done in Arjovsky et al. (2017)). Pseudocode for the proposed method is presented in Algorithm 2.

Algorithm 2 Algorithm for proposed method *RWIGAN*

Input: Data matrix $\mathbf{X}_{NA}$ with missing entries, and $\mathbf{M}$ (implicit in $\mathbf{X}_{NA}$)

Initialize:

- Randomly initialize Θ_G, Θ_D using the *Xavier* Initialization (Yoon et al. (2018a)).
- Standardize every (observed) data column to be between 0 and 1

while $\nabla_{\Theta_G}\mathcal{L}_G^t$ at iteration t does not converge **do**

Sample batch with indices $\mathbf{B}$. For each step assume all data (i.e. $\mathbf{X}, \mathbf{M}, \dot{\mathbf{X}}, \dot{\mathbf{M}}$) are subset at $\mathbf{B}$:

- Generate random uniform(0,1) matrix $\mathbf{U}$ (subset at $\mathbf{B}$)
- $\mathbf{X} \leftarrow \mathbf{X}_{NA} \odot \mathbf{M} + (1 - \mathbf{M}) \odot \mathbf{U}$, then create $\dot{\mathbf{X}}$ using generator G: $\dot{\mathbf{X}} := G(\mathbf{X}|\Theta_G)$.
- (1) "D-Optimization": update Θ_D by minimizing the D loss function using SGD: $\nabla_{\Theta_D}\mathcal{L}_D^{\lambda}(\mathbf{X}, \mathbf{M}, \Theta_D)$
  - $\hat{\mathbf{X}} \leftarrow \mathbf{X} \odot \mathbf{M} + (1 - \mathbf{M}) \odot \dot{\mathbf{X}}$
  - $\hat{\mathbf{M}} \leftarrow D(\hat{\mathbf{X}}|\Theta_D)$.
- (2) G-optimization: update Θ_G by SGD using $\nabla_{\Theta_G}\mathcal{W}(P_{\text{data}}, P_G, \Theta_D)$ using the inverse quantile re-calibration in Algorithm 1.

end while

4 Simulation Study and Method Comparison

We design and conduct experiments on several sets of synthetic data. We simulate two types of experiments: (1) *informative missing* (IM), where the missingness of a variable is dependent on itself, (2) covariate imbalance (CI), where the missingness of variable is dependent on other variables (in addition to IM). The proposed method is primarily designated for heavy-tailed distributions whose patterns of missingness are self-dependent. However, we model the missingness as dependent on both latent and observed patterns. Most DL methods are not catered to MNAR data (Yoon et al. (2018a); Wang et al. (2021)), except under very specific conditions Ma and Zhang (2021); Dai et al. (2021). Albers et al. (2018) note that real EHR datasets are rarely comprised of multivariate normal variables. In the following Section 5 this phenomena is evident in REACH data, so we simulate several scenarios involving heavy-tailed distributions that look similar to the distributions of the EHR data that is used in the case studies. We simulate several high-dimensional datasets with IM (details in Table 1, right). Variables such as glucose, lipase, and creatinine are commonly modeled as lognormal or generalized extreme value (GEV) distributions (Albers et al. (2018)).

EHR data in real-world examples roughly look like the simulations in this section. We simulate several cases of data with IM and CI. Each of these distributions have 300 observations and 100 variables. In each simulation, we generate either correlated or independent random variables or mixtures of random variables. We generate independent exponential (.1) and uniform variables for the IM experiments. Higher values are more likely to be missing as to be *informative* in the experiments. Details of missingness are described in Table 1 (right). Further details can be found in Appendix E.2. These modes of missingness partially match those from the PCL variables from REACH data in Section 5. In IM-exp, each variable is additionally missing if the value of the entry is greater than the 80% quantile of the variable. In the IMCI cases, variables were generated in pairs (Y_j, Z_j), such that the final dataset $\mathbf{X}$ is their concatenation . but the variables are correlated with each othe. Details of these simulations are described in Table 1. In order to ease interpretation and avoid skewed results (Gondara and Wang (2017)), all columns were scaled between 0 and 1.

4.1 Method Comparison

To evaluate the proposed method, we designed experiments on synthetic datasets and compared it with MICE, MIDA, and GAIN. MICE, as we have described, is considered state of the art for missing data. GAIN is theoretically sound and has partly inspired the framework of RWIGAN, but it does have limitations: Wang et al. (2021) have shown that it does not impute as well as MICE. We also compare to another method called *multiple imputation using denoising autoencoders* (MIDA) (Gondara and Wang (2017)). MIDA uses a *decoder*, which outputs a low-dimensional approximation of the input data, and an *encoder* which projects the approximation to the back to the data space. This two step process is similar to that of GAIN, but does not rely on the missingness structure. We simulate 50 Monte Carlo replicates for each distributional specification. Operationally, because MICE is much slower than the other methods, we only use 20 replicates. MSE and normalized MSE (nMSE) (details in Appendix E.1), where each variable is scaled by the (true) column mean, were used to assess the model performance (Yoon et al. (2018a); Gondara and Wang (2017)).

Mean Squared Error (MSE)

Label	*Distribution*	RWIGAN	GAIN	MIDA	MICE
IM-exp	exp(.1)	.04± .001	.07± .002	.05± .001	.05±.006
IM-u-a	unif (0,1)	.40±.015	.38±.002	.45±.008	.49± .008
IM-u-b	unif(0,1)	.09± .003	.27±.012	.20± .004	.19 ±.006
IMCI-a	MVN(.25), exp(.5)	.09±.003	.28±.009	.16± .002	.14±.002
IMCI-b	exp(.1), exp(.01)	.039±.002	.061±.004	.047±.002	.045±.003

Normalized Mean Squared Error (NMSE)

Label	*Distribution*	RWIGAN	GAIN	MIDA	MICE
IM-exp	exp(.01)	.19 ± .005	.32±.003	.26±.007	.24± .005
IM-u-a	unif(0,1)	.32 ±.003	.33±.004	.46±.018	.55±.015
IM-u-b	unif(0,1)	.13± .003	.44±.012	.32±.004	.28± .010
IMCI-a	MVN(.25), exp(.5)	.16 ± .007	.54 ±.020	.29 ±.007	.24± .004
IMCI-b	exp(.1), exp(.01)	.19 ±.005	.31 ±.014	.24±.010	.21±.007

Legend

	Distribution	Missingness
IM-exp	$X_j \sim \exp(.1)$ $\mathbf{X} = \{X_j\}_p$	$\mathbb{P}_{\text{msg}}(X_{ij}) = \frac{\log(X_{ij}+1)}{\max_i(\log(X_{ij}+1))}$, $\mathbb{P}_{\text{msg}}(X_{ij}) = 1$ if $X_{ij} \geq$80% qtl
IM-u-a	$X_j \sim$ unif(0,1) $\mathbf{X} = \{X_j\}_p$	$\mathbb{P}_{\text{msg}}(X_{ij}) = .50$ if $X_{ij} \geq 70\%$ qtl
IM-u-b	$X_j \sim$ unif(0,1) $\mathbf{X} = \{X_j\}_p$	Same as (IM-exp)
IMCI-a	$U_j \sim$ unif(0,1) $\mathbf{Z} \sim$ MVN(0, 1, $\rho = .25$) $Y_j \sim 1+2U_j+X_j + \exp(.5)$ $\mathbf{X} = \{Y_j, Z_j\}_p$	$\mathbb{P}_{\text{msg}}(Z_j)$ as IM-exp $\mathbb{P}_{\text{msg}}(Y_j, Z_j) \propto$ only Z_j U_j is $\perp$ latent variable
IMCI-b	$Z_j \sim \exp(.01)$ $Y_j \sim 1 + 2X_j + \exp(.1)$ $\mathbf{X} = \{Y_j, Z_j\}_p$	$\mathbb{P}_{\text{msg}}(Z_j)$ as IM-exp $\mathbb{P}_{\text{msg}}(Y_j, Z_j) \propto$ Z_j only

Table 1: Simulations for *Informative Missing* (IM), and IM with *Covariate Imbalance* (IMCI). All variables except for the (standard) multivariate normals are independent. The right side is a key describing the $\mathbf{X}$ represents the total dataset for imputation. $\mathbb{P}_{\text{msg}}(X_j)$ represents the missingness rates of the $j-$th variable X_j. In each scenario the number of observations $n =$ is 300 and variables $p = 100$. The simulations are normalized between 0 and 1. For each IMCI experiment, 50 X_j and Y_j are generated and the resulting matrix $\mathbf{X}$ is comprised from their concatenation. MSE and nMSE of the imputed values versus the ground-truth values of the induced-missing values were calculated as a measure of performance. 50 synthetic experiments for each setting is carried out and the Monte Carlo average of the MSE and nMSE are presented.

RWIGAN outperforms other methods in most cases. For IM-u-a (where the missingness structure is more simplistic and closer to MAR), RWIGAN outperforms most other methods except for GAIN (for MSE). Though our proposed method works across a range of modes of missingness, the closer the missingness is to randomness, the less obvious its advantage is in relation to other methods. Figure 2 shows one case of the IM simulations from Table 1 and its imputations by the different methods. The reconstruction contour density plots from RWIGAN (left) is centered around the median for both true and imputed values; they are much less skewed than in the other methods. In terms of speed, on average, MICE is the slowest and expends about 600 CPU seconds per run (for REACH it takes 1000, and is prohibitively slow for INSIGHT). Empirically, a simulation dataset of $n = 300, p = 100$ takes MICE approximately 15 minutes. DL methods (RWIGAN, GAIN, and MIDA) are much faster, especially when p becomes large. RWIGAN is slower than MIDA, which is in turn slower than GAIN, but the increase in time with respect to increased p or n is linear.

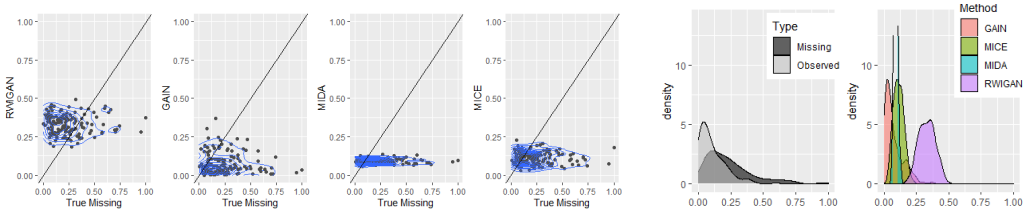

Figure 2: Imputation results compared to the true values (left) for different methods for the IM simulation and densities of missing and observed entries and reconstructed densities (right).

5 Case Study

We use two sources of Electronic Health Records: (1) REactions to Acute Care and Hospitalization (REACH) and (2) INSIGHT clinical research network. We develop a method of evaluating the missingness in imputations by modeling the existing missingness and then to induce *additional* missingness in the observed parts of the variables. REACH is a rich dataset that contains the mental and physical health conditions of 1776 subjects from Columbia University Medical Center. These

EHR include symptoms, lab results, and other related information. A major function of REACH is to analyze the association between post-traumatic stress disorder (PTSD) and cardiovascular risk. The PTSD Checklist (PCL) is a continuous, long-tailed metric and has shown "convergent and discriminant validity, [internal] consistency, and [test-test] reliability" Ruggiero et al. (2003). Indeed, many subsequent measures focus on the various aspects of the PCL score. In many cases, PCL scores function as the predicted variables and is the driving factor of the analysis (Birk et al. (2019)). The INSIGHT Clinical Research Network houses EHR data of over 12 million patients in New York City from its largest private healthcare systems (INS) .

REACH yields two overlapping datasets: a *raw* where $n = 1776$, and *cleaned* where $n = 764$. The *raw* dataset has more missing, but some of these values are observed in the *cleaned* (more details in Table 3). We conduct two analyses on both the datasets; for the *cleaned* analysis we create NAs in the observed data where the *raw* data is missing, and use the real values as ground truth. Exploratory analysis (Figure 7 in Appendix) shows that the densities between the observed and missing look inherently different; details of missingness modeling can be found in Appendix D.1. For the (larger) *raw* analysis, we simulate artificial NAs because we do not know the ground truth; we "bootstrap" a notion of ground truth for missing data in the *raw analysis* and model the missingness of the *raw* variables by using the difference in missingness between the *raw* and *cleaned* datasets. We follow prior studies in emphasizing the imputation on PCL variables Birk et al. (2019). Results show that RWIGAN is preferable to GAIN and MIDA consistently in terms of MSE, nMSE, as well as the (canonical) W_1 (1) in both *raw* and *cleaned* cases. MICE does slightly better than RWIGAN in some cases, but it is much slower. In the smaller analysis, the advantages of RWIGAN is less pronounced because the real rates of missingness are not as strongly defined as the induced missingness in the largest, but the completeness of REACH may not be indicative of typical EHR.

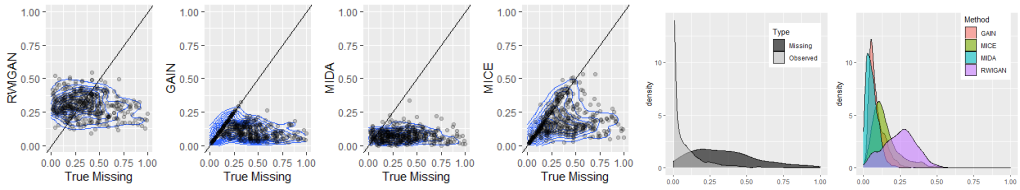

Figure 3: (Left) Contour plots of imputed values for the REACH PCL Combo (Baseline) data. The MSE for each method for this variable in particular is QWIGAN: .066, GAIN: .115, MIDA: .130, MICE: .079. (Right) density plots are shown for the observed, missing, and imputed values

RWIGAN's advantages are apparent in both the *raw* and *cleanred* analyses. In the full data, the nMSE, MSE, and average W_1 (Table 2, third row). In the smaller analysis with real missingness, for PCL variables, RWIGAN results yields the lowest nMSE, and comparatively low errors across all measures. Figure 3 shows an example of resultant imputations plotted against induced missing values for the PCL (baseline) variable. Though the variance of imputations among is fairly wide, RWIGAN suffers from much less bias than the other methods. MICE does well in minimizing errors (Table 2), but the numerical results only show one side of the story: visual representations in Figure 3 show some of the more nuanced advantages of our proposed method.

The INSIGHT dataset has many different datasets such as diagnoses, vital statistics, and laboratory tests. We center this analysis on the laboratory tests. Albers et al. (2018) emphasize the importance of lab tests, but also describe them as "noisy, outlier-ridden,and biased". We aggregated quantitative results of all lab tests for every patient who tested positive for COVID-19 from 2020-2021 (with some thresholds). Details of the data and missingness are described in Table 2; further specificities are in Appendix D.3. The resultant data is fairly sparse; each variable is heavy tailed. This is indicative of the nature of most EHR data; Figure 4 shows a few of these variables. Unlike REACH, INSIGHT does not have any internal benchmarks for assessing missingness; we induce NAs on the observed variables with the same scheme that was used in REACH. Table 2 shows that RWIGAN outperforms the other methods across a several metrics for INSIGHT. MICE was not used because it was prohibitively slow. Figure 4 show imputation comparisons for a single variable; competing methods again produce more bias.

Across simulations and data, our proposed method reconstructs missing data more effectively both numerically and visually. Density plots show that RWIGAN imputations are more similar to the

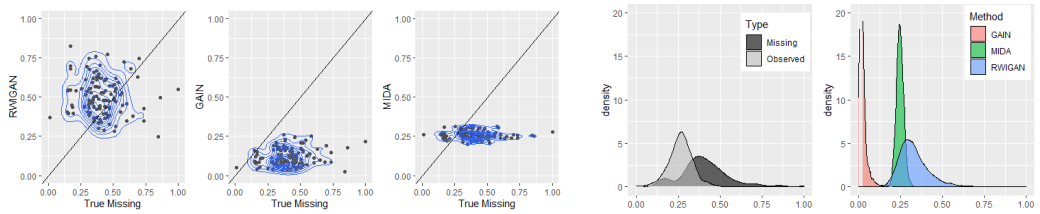

Figure 4: Density and contour plots for various imputations of *Carbon dioxide (partial pressure) in blood* variable from *Insight*. Left shows contour plots for the (induced) missing vs imputed values for each of the methods, right shows density plots of the true observed and missing values and imputations.

original data; it does not look exactly like the *observed* nor *missing* densities but a mixture. We use MSE and W_1 as the tools to judge the consensus imputation results but these are not perfect tools. Point-by-point inspections of distributions in Figures 3, and 4 show that RWIGAN produces more less unbiased and centered imputations. In GAIN, a "hint" (subsample of $\mathbf{M}$) is used in training to encourage distributional similarity between missingness and observed. In RWIGAN we omit this and note instead that the $\mathcal{W}$ recalibration achieves the same goal of distributional equivariance, if not moreso empirically as shown by Figures 2-4. Additional visual examples (more typical of GAN studies) are found in Appendix E.3.

Method Evaluation for Real EHR Data

		REACH				**INSIGHT**			
Var.	*Metric*	RWIGAN	GAIN	MIDA	MICE	*Metric*	RWIGAN	GAIN	MIDA
PCL	nMSE	.47±.020	.63±.020	.76±007	.48±.005	nMSE	.117 ±.001	.315 ±.010	.132 ±.006
(full)	MSE	.068±.004	.089±.004	.10±.001	.067±.002	MSE	.005 ±.001	.020 ±.002	.006 ±.002
	$\bar{W}_1$	.18±.002	.23±.006	.30 ±.003	.20±.005	$\bar{W}_1$	.116 ±.006	.310 ±.005	.171±.003
PCL	nMSE	1.1± .051	2.2±.056	1.2±.120	1.4±.071				
(small)	MSE	.08±.053	.19±.004	.06±.004	.09±.009				
	$\bar{W}_1$	.57±.076	.79±.017	.65±.017	.49±.047				

Legend

Label	Missingness	(n, p)(eval.)	(n, p) (analysis)
PCL (full)	$\mathbb{P}_{\text{msg}}(X_j) = \log(\kappa_j + 1)/\max(\log(\kappa_j + 1))$ where $\kappa_j = (n_j/4 - \text{rank}(X_j))^2/n_j$	(1776,24)	(1776,62)
PCL (small)	Real missingness	(764,4)	(764,13)
INSIGHT	(Same as PCL(full))	(4764,162)	(4764,162)

Table 2: Imputation evaluations for MSE and nMSE for REACH and INSIGHT data. Each variable is scaled between 0 and 1. Results on top left is for the full REACH data and bottom left is for the *cleaned* data only. Only PCL variables are evaluated for REACH. Results for INSIGHT are in the top right; all variables are used and evaluated. MICE is not used for INSIGHT because it is prohibitively slow.

6 Discussion

RWIGAN is a novel algorithm to impute missing data when the missingness is not at random. The efficacy of the method is supported by the results of synthetic experiments (Section 4.1) and real data (Section 5). Our proposed method generally outperforms existing methods for real data. MICE does slightly better in some cases but it could not computationally handle the sizes of larger datasets like INSIGHT. Another contribution of this work in DL-based imputation studies is in our usage of high-quality real clinical data; in REACH we have a quasi-experimental setting for assessing benchmarks of missing data where certain data are missing at one time point but observed at another. RWIGAN is designed for specific forms of EHR data, which is heavy tailed and likely to be informative missing (Albers et al. (2018)). We have demonstrated empirical evidence for IM as a real mechanism of missingness in EHR in Appendix D.1. We discussed in Section 2 that existing WGAN methods do not actually serve as W_1 minimizers (Stanczuk et al. (2021)), which leaves ample room for redesigning criterion (i.e. recalibration) for DL-based imputation. Ostrovski et al. (2018) have investigated the relationship between quantiles and Wasserstein distance, but further work will be fruitful. We have demonstrated the empirical advantages of the recalibrated Wasserstein metric $\mathcal{W}$, further work may explore its theoretical properties.

REFERENCES

Insight clinical research network. URL https://its.weill.cornell.edu/services/research-informatics/insight-clinical-research-network.

D.J. Albers, N. Elhadad, J. Claassen, R. Perotte, A. Goldstein, and G. Hripcsak. Estimating summary statistics for electronic health record laboratory data for use in high-throughput phenotyping algorithms. *Journal of Biomedical Informatics*, 78:87–101, 2018. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2018.01.004. URL https://www.sciencedirect.com/science/article/pii/S1532046418300066.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN, 2017. URL https://arxiv.org/abs/1701.07875.

Jeffrey Birk, Ian Kronish, Bernard Chang, Talea Cornelius, Marwah Abdalla, Joseph Schwartz, Joan Duer-Hefele, Alexandra Sullivan, and Donald Edmondson*. The Impact of Cardiac-induced Posttraumatic Stress Disorder Symptoms on Cardiovascular Outcomes: Design and Rationale of the Prospective Observational Reactions to Acute Care and Hospitalizations (ReACH) Study. *Health Psychol Bull.*, 3(1):10–20, January 2019. doi: 110.5334/hpb.16. URL https://www.hal.inserm.fr/inserm-02310563.

L. Breiman, Jerome H. Friedman, Richard A. Olshen, and C. J. Stone. Classification and regression trees. 1983.

Hao Cheng and Ying Wei. A fast imputation algorithm in quantile regression. *Comput. Stat.*, 33 (4):1589–1603, dec 2018. ISSN 0943-4062. doi: 10.1007/s00180-018-0813-z. URL https://doi.org/10.1007/s00180-018-0813-z.

Zongyu Dai, Zhiqi Bu, and Qi Long. Multiple imputation via generative adversarial network for high-dimensional blockwise missing value problems. *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 791–798, 2021.

William Fedus, Mihaela Rosca, Balaji Lakshminarayanan, Andrew M. Dai, Shakir Mohamed, and Ian Goodfellow. Many paths to equilibrium: Gans do not need to decrease a divergence at every step, 2017. URL https://arxiv.org/abs/1710.08446.

Magda Friedjungová, Daniel Vašata, Maksym Balatsko, and Marcel Jiřina. Missing features reconstruction using a wasserstein generative adversarial imputation network. In Valeria V. Krzhizhanovskaya, Gábor Závodszky, Michael H. Lees, Jack J. Dongarra, Peter M. A. Sloot, Sérgio Brissos, and João Teixeira, editors, *Computational Science – ICCS 2020*, pages 225–239, Cham, 2020. Springer International Publishing. ISBN 978-3-030-50423-6.

Lovedeep Gondara and Ke Wang. Mida: Multiple imputation using denoising autoencoders, 2017. URL https://arxiv.org/abs/1705.02737.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

John W. Graham. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60(1):549–576, 2009. doi: 10.1146/annurev.psych.58.110405.085530. URL https://doi.org/10.1146/annurev.psych.58.110405.085530. PMID: 18652544.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein GANs, 2017. URL https://arxiv.org/abs/1704.00028.

Jerry Li, Aleksander Madry, John Peebles, and Ludwig Schmidt. On the limitations of first-order approximation in GAN dynamics, 2017. URL https://arxiv.org/abs/1706.09884.

Steven Cheng-Xian Li, Bo Jiang, and Benjamin Marlin. Misgan: Learning from incomplete data with generative adversarial networks, 2019. URL https://arxiv.org/abs/1902.09599.

R.J. Little and D.B. Rubin. *Statistical Analysis with Missing Data*.

Yonghong Luo, Xiangrui Cai, Ying ZHANG, Jun Xu, and Yuan xiaojie. Multivariate time series imputation with generative adversarial networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/96b9bff013acedfb1d140579e2fbeb63-Paper.pdf.

Chao Ma and Cheng Zhang. Identifiable generative models for missing not at random data imputation. *CoRR*, abs/2110.14708, 2021. URL https://arxiv.org/abs/2110.14708.

Georg Ostrovski, Will Dabney, and Rémi Munos. Autoregressive quantile networks for generative modeling, 2018. URL https://arxiv.org/abs/1806.05575.

Kenneth J. Ruggiero, Kevin Del Ben, Joseph R. Scotti, and Aline E. Rabalais. Psychometric properties of the ptsd checklist—civilian version. *Journal of Traumatic Stress*, 16(5):495–502, 2003. doi: https://doi.org/10.1023/A:1025714729117. URL https://onlinelibrary.wiley.com/doi/abs/10.1023/A%3A1025714729117.

Shaun R Seaman and Ian R White. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295, 2013. doi: 10.1177/0962280210395740. URL https://doi.org/10.1177/0962280210395740. PMID: 21220355.

Luciana O. Silva and Luis E. Zárate. A brief review of the main approaches for treatment of missing data. *Intell. Data Anal.*, 18(6):1177–1198, nov 2014. ISSN 1088-467X.

Jan Stanczuk, Christian Etmann, Lisa Maria Kreusser, and Carola-Bibiane Schönlieb. Wasserstein gans work because they fail (to approximate the wasserstein distance), 2021. URL https://arxiv.org/abs/2103.01678.

Laszlo Talas, John G. Fennell, Karin Kjernsmo, Innes C. Cuthill, Nicholas E. Scott-Samuel, and Roland J. Baddeley. Camogan: Evolving optimum camouflage with generative adversarial networks. *Methods in Ecology and Evolution*, 11(2):240–247, 2020. doi: https://doi.org/10.1111/2041-210X.13334. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13334.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03. URL https://www.jstatsoft.org/index.php/jss/article/view/v045i03.

Cédric Villani. Optimal transport: Old and new. 2008.

Zhenhua Wang, Olanrewaju Akande, Jason Poulos, and Fan Li. Are deep learning models superior for missing data imputation in large surveys? evidence from an empirical comparison, 2021.

Ying Wei and Yunwen Yang. Quantile regression with covariates missing at random. *Statistica Sinica*, 2014.

Ying Wei, Yanyuan Ma, and Raymond J. Carroll. Multiple imputation in quantile regression. *Biometrika*, 99(2):423–438, 2012. URL https://ideas.repec.org/a/oup/biomet/v99y2012i2p423-438.html.

Yanmei Xie and Biao Zhang. Empirical likelihood in nonignorable covariate-missing data problems. *The International Journal of Biostatistics*, 13, 01 2017. doi: 10.1515/ijb-2016-0053.

Yinchong Yang, Zhiliang Wu, Volker Tresp, and Peter A. Fasching. Categorical EHR imputation with generative adversarial nets. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, jun 2019. doi: 10.1109/ichi.2019.8904717. URL https://doi.org/10.1109%2Fichi.2019.8904717.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Gain: Missing data imputation using generative adversarial nets, 2018a. URL https://arxiv.org/abs/1806.02920.

Jinsung Yoon, William R. Zame, Amitava Banerjee, Martin Cadeiras, Ahmed M. Alaa, and Mihaela van der Schaar. Personalized survival predictions via trees of predictors: An application to cardiac transplantation. *PLOS ONE*, 13(3):1–19, 03 2018b. doi: 10.1371/journal.pone.0194985. URL https://doi.org/10.1371/journal.pone.0194985.

Seongwook Yoon and Sanghoon Sull. Gamin: Generative adversarial multiple imputation network for highly missing data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8453–8461, 2020. doi: 10.1109/CVPR42600.2020.00848.

Xiaohui Yuan and Xiaogang Dong. Weighted empirical likelihood for quantile regression with non ignorable missing covariates. *Communications in Statistics - Theory and Methods*, 48(12):3068–3084, 2019. doi: 10.1080/03610926.2018.1473604. URL https://doi.org/10.1080/03610926.2018.1473604.

AUTHOR CONTRIBUTIONS

Omitted for blind review

ACKNOWLEDGMENTS

Omitted for blind review

A WASSERSTEIN DISTANCE

We first define Wasserstein distance. Let $\mathcal{X}$ be a Borel set with $\mathbb{P}(\mathcal{X})$ as the set of probability measures defined on $\mathcal{X}$. For two distributions $P_S, P_S \in \mathbb{P}(\mathcal{X})$, $\Pi(P_R, P_S)$ is the set of all possible *joint* distributions $\gamma(x, y)$ (x, y are elements of $\mathcal{X}$) whose marginal distributions are P_R, P_S. $\gamma(x, y)$ indicates how much mass must be transported from x to y in order to transform the distribution P_R *into* P_S. The Wasserstein-1 distance is the cost of the *optimal transport* plan.

$$W_1(P_R, P_S) = \inf_{\gamma \in \Pi(P_R, P_S)} \mathbb{E}_{(x,y)\sim\gamma(x,y)} \|x - y\| \tag{8}$$

Let $\|x - y\|$ as the analogous to *cost*, of, for example, of moving bread from bakery x to cafe y Villani (2008). Under the first definition of Wasserstein Distance, the objective is to minimize the cost of transporting bread.

Perhaps the most important consequence of Arjovsky et al. (2017) is their application of Kantorovich-Rubinstein Duality to the minimax optimization problem in Goodfellow et al. (2014). The duality is a re-conceptualization of the cost-minimization problem in production to a price-maximization problem in distribution. Suppose the ownership of the bakery and cafe is controlled by a collective that produces as well as sells goods. $f(\cdot)$ serves as the analogy for the *price* that the collective sets after the bread is baked. $f(x)$ represents the price that the bread is bought at barkery x, and $f(y)$ is the price at which it is sold at cafe y. Now the whole cost that the collective pays is $f(y) - f(x)$ for the total cost of transport as the total revenue instead of simply $|y - x|$. f is Lipschitz because of the material constraints in transforming the dough into a sellable good (i.e. bread is only as good as the raw materials that make it). To be competitive, the bakery-cafe collective must set up prices in a way that $f(y) - f(x) \leq |y - x| \quad \forall x, y$ Then, the duality comes from when the collective makes decisions instead of only the bakery: the bakery seeks to minimize costs, but the collective seeks to maximize the profits while keeping the costs constrained at a minimum. Mathematically, this can be written as:

$$\sup \Big\{ \int_{\mathcal{X}} f(y) \mathrm{d}P_R(y) - \int_{\mathcal{X}} f(x) \mathrm{d}P_S(x); \quad f(y) - f(x) \leq |y - x| \Big\}$$

The Kantorovich Duality states that an equivalent definition of the Wasserstein distance is the supremum of differences in expectations of densities with Lipschitz functions f.

$$W_1(P_R, P_S) \equiv \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim P_R}[f(x)] - \mathbb{E}_{y\sim P_S}[f(y)]$$

the supremum of the Lipschitz functions f is analogous to the maximization of the discriminator. Let P_{data} denote the data distribution , P_G is distribution of the generator output . The global objective is to minimize the Wasserstein distance between the generated and the 'real' distribution:

$$\min_{\Theta_G} W_1(P_{\text{data}}, P_G) = \min_{\Theta_G} \max_{\Theta_D : D \in \mathcal{D}} \left(\mathbb{E}_{\mathbf{X}\sim P_{\text{data}}(\mathbf{X})} D(\mathbf{X}) - \mathbb{E}_{\dot{\mathbf{X}}\sim P_G}[(D(\dot{\mathbf{X}})] \right)$$

where $\mathcal{D}$ is set of all 1-Lipschitz functions. WGAN Arjovsky et al. (2017) enforces the Wasserstein distance by clipping the estimated weights, which has many problems such as exploding and vanishing gradients. Guljarani et al discuss improved training of Wasserstein GANS using the *gradient penalty* Gulrajani et al. (2017).

B EQUIVALENCE UNDER INFORMATIVE MISSING THROUGH PROBABILITY REDISTRIBUTION

Under informative missingness, the distributions of the observed data and missing data are directly not equivalent, i.e. $P_{\text{obs}} \neq P_{\text{miss}}$. However, one could re-acheive the distributional equivalence by re-distributing the probability mass of P_{obs} and P_{msg}. To illustrate this concept, let's consider a simple case without covariates, where $\mathbb{P}(\delta_i = 1) = \pi(y_i)$. Let δ be the missingness vector of a single vector (i.e. M_j for some variable j). Following Bayes theorem, we have

$$P(Y \mid \delta = 1) \propto P(\delta = 1 \mid y)p(Y), \text{ and } P(Y \mid \delta = 0) \propto P(\delta = 0 \mid y)p(Y).$$

It follows that

$$\frac{P(Y \mid \delta = 1)}{\pi(Y)} \propto p(Y) \text{ and } \frac{P(Y \mid \delta = 0)}{1 - \pi(Y)} \propto p(Y) \tag{9}$$

In other words, if we adjust the distribution of observed data (i.e. $P(Y \mid \delta = 1)$) by $\pi(Y)$ and adjust the distribution of missing data (i.e. $P(Y \mid \delta = 0)$) by $1 - \pi(Y)$, the two adjusted distributions should be equivalent. Such distributional equivalence is the cornerstone of the proposed algorithm.

Let n_0 be the number of missing $y_i's$, and $n_1 = n - n_0$ is the number of observed y_i's. The empirical distributions of the observed $y_i's$ and missing $y_i's$ can be written as

$$\widehat{F}_n^{(obs)}(x) = \sum_{i=1}^{n} \frac{1}{n_1} \mathbf{1}\{y_i < x\}\mathbf{1}\{\delta_i = 1\}, \text{ and } \widehat{F}_n^{(msg)}(x) = \sum_{i=1}^{n} \frac{1}{n_0} \mathbf{1}\{y_i < x\}\mathbf{1}\{\delta_i = 0\}.$$

In other words, each of observed y_i in P_{obs} receives a probability mass of $1/n_1$, and each of missing y_i in P_{mis} receives a probability mass of $1/n_0$. Following (9), we re-assign the probability mass of the observed y_i and the missing y_js as in the following table

		Observed data (y_i's)	Missing data (y_j's)
Probability mass on	Original	$\frac{1}{n_1}$	$\frac{1}{n_0}$
individual observations	Reassigned	$\omega_i = \frac{1/\pi(y_i)}{\sum_{k=1}^{n_1} 1/\pi(y_k)}$	$\omega_j = \frac{1/(1-\pi(y_j)}{\sum_{k=1}^{n_0} 1/(1-\pi(y_k))}$

Replacing the original equal weights by ω_i and ω_j, we reconstruct empirical distributions by

$$\widetilde{F}_n^{(obs)}(x, \omega) = \sum_{i=1}^{n_1} \omega_i \mathbf{1}\{y_i < x\}, \text{ and } \widetilde{F}_n^{(msg)}(x, \omega) = \sum_{j=1}^{n} \omega_j \mathbf{1}\{y_j < x\}.$$

Following the Bayesian theorem, we expect the re-weighted distributions $\widetilde{F}_n^{(obs)}(x, \omega)$ and $\widetilde{F}_n^{(msg)}(x, \omega)$ are equivalent. Such distributional equivalance is the corner stone of the proposed algorithm. In the proposed GAN framework, we use Discriminator D to generate the weights ω_i and use Generator G to generate imputations, and interatively updates D and G to achieve a global optimization in re-weighted distributional equivalence.

Sample Illustration To illustrate this idea, we generate a random sample (y_i, δ_i), where y_i follows a $u(0, 1)$ distributions, and δ_i is a binomial distribution with $\mathbb{P}(\delta_i = 1) = \log(y_i + 1)$. As result, the higher the value of y_i, the less likely to be missing. Overall, it leads to approximately 62% y_i's are missing. In figure 1 (a), we illustrate the empirical distributions of the observed y_i's (i.e. those $\delta_i = 1$) and "missing" y_i's (i.e. those $\delta_i = 0$). It is clear that they are different from each other.

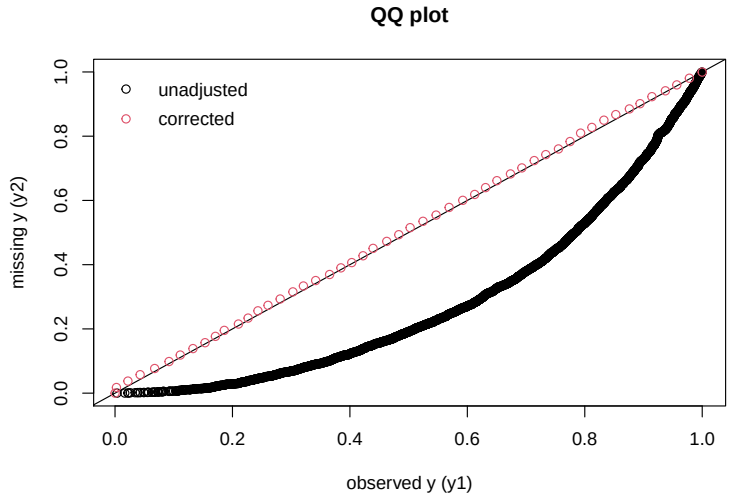

Figure 5: QQplot of the distributions of observed and missing y's under informative missing before and after corrections.

Quantile function is an equivalent form of distribution function. The re-weighted distributional equivalence also indicates the quantile equivalence. Let $\{y_{(1)}^{(obs)}, y_{(2)}^{(obs)}, ..., y_{(n_1)}^{(obs)}\}$ be the order statistics of the observed y_i's, and $(\omega_{(1)}, ..., \omega_{(i)}, ..., \omega_{(n_1)})$ are their corresponding re-assigned probability mass; Likewise, we define $\{y_{(1)}^{(mis)}, y_{(2)}^{(mis)}, ..., y_{(n_0)}^{mis)}\}$ as the order statistics of the missing y_j's, and $(\omega_{(1)}, ..., \omega_{(j)}, ..., \omega_{(n_0)})$ are their corresponding re-assigned probability masses. The re-weighted empirical quantile functions can be written as

$$\widetilde{Q}_n^{(obs)}(\tau, \omega) = \inf\{y_{(k)}^{(obs)} : \sum_{i=1}^{k} \omega_{(i)} \geq \tau;\ k = 1, ..., n_1\},\ \text{and}$$

$$\widetilde{Q}_n^{(mis)}(\tau, \omega) = \inf\{y_{(k)}^{(mis)} : \sum_{j=1}^{k} \omega_{(j)} \geq \tau;\ k = 1, ..., n_0\}$$

We expect $\widetilde{Q}_n^{(obs)}(\tau, \omega)$ and $\widetilde{Q}_n^{(obs)}(\tau, \omega)$ to be equivalent as well, and use a recalibrated Wasserstein Distance to assess quantile equivalence in the proposed RWIGAN imputation algorithm.

C Details of the Proposed Method

C.1 Neural Network Architecture

G and D are both fully connected 3-layer neural networks. The input for G is: $\mathbf{X}$ with dimensions $n \times p$.

$$h_1^G = \text{relu}(W_G^1 \mathbf{X} + b_G^1)$$
$$h_2^G = \text{relu}(W_G^2 h_1^G + b_G^2)$$
$$\dot{\mathbf{X}} = \text{relu}(W_G^3 h_2^G + b_G^3)$$

The output is the imputation $\dot{\mathbf{X}}$, also with dimensions $n \times p$. We use the relu activation function in this case but the elu or sigmoid functions may also be used. The input for D is: $\mathbf{X}$ (or $\dot{\mathbf{X}}$, or $\hat{\mathbf{X}}$,) with dimensions $n \times p$.

$$h_1^D = \text{relu}(W_D^1 \mathbf{X} + b_D^1)$$
$$h_2^D = \text{relu}(W_D^2 h_1^D + b_D^2)$$
$$\tilde{\mathbf{M}} = \text{relu}(W_D^3 h_2^D + b_D^3)$$

The output is a probability of the mask matrix $\tilde{\mathbf{M}}$, as in GAIN Yoon et al. (2018a).

C.2 CRITERION OF G

The criterion function of the generator G, in contrast with the discriminator, is the following minimization.

$$\min_{\Theta_G} \mathcal{L}_G(\mathbf{X}, \mathbf{M}, \Theta_G) = \min_{\Theta_G} \mathbb{E}_{\hat{\mathbf{X}} \sim P_{\text{data}}}[\mathbf{X}|\mathbf{M}] - \mathbb{E}_{G(\mathbf{X}) \sim P_G}[D(G(\mathbf{X}, \Theta_G))|\mathbf{M}]. \tag{10}$$

by the Kantorovich-Rubinstein Duality (which motivates Wasserstein GAN in the first place), the G likelihood is equivalent to the Wasserstein distance

$$\begin{aligned}
\mathcal{L}_G(\mathbf{X}, \mathbf{M}) &= \sup_{\Theta_D \in \mathcal{D}} \mathbb{E}_{\mathbf{X} \sim P_{\text{data}}}[D(\mathbf{X})|\mathbf{M}] - \mathbb{E}_{\dot{\mathbf{X}} \sim P_G}[D(\dot{\mathbf{X}})|\mathbf{M}] \\
&= \sup_{\|f\|_L \leq 1} \mathbb{E}_{\mathbf{X} \sim P_{\text{data}}}[f(\mathbf{X})] - \mathbb{E}_{\dot{\mathbf{X}} \sim P_G}[f(\dot{\mathbf{X}})] \\
&= \mathcal{W}(P_G, P_{\text{data}}).
\end{aligned}$$

where, in this case, the function $f(\cdot)$ represents the conditional expectation $\mathbb{E}[D(\cdot)|\mathbf{M}]$ of the discriminator given mask M.

C.3 GENERATOR LOSS IS DIFFERENTIABLE AND MONOTONE

Recall that $\hat{\mathbf{M}}$ is the estimated mask of $\hat{\mathbf{X}}$. Since in G-minimization the Θ_D is fixed, we write $\hat{M}_X(\Theta_G)$ only as a function of Θ_G. Let M be a *mask vector* (i.e. M_j for some arbitrary variable index j). WLOG let X be a data vector, also with the index j omitted. To simplify notation, we define $\hat{M}_X(\Theta_G)$ as the output probability from D, depending on Θ_G for backpropagation.

$$\begin{aligned}
\hat{M}_X(\Theta_G) &:= D(\hat{\mathbf{X}}|\Theta_D)_j \\
&= D(\mathbf{M}_{(j)} \odot \mathbf{X}_{(j)} + (1 - \mathbf{M}) \odot G(\mathbf{X}, \Theta_G)_{(j)}|\Theta_D).
\end{aligned}$$

for some arbitrary j. Note that $\mathbf{\Phi}_X(x, \hat{M}_X|M_i = 1)$ does not depend on G so we omit for now. But $\mathbf{\Phi}_{\dot{X}}(x, 1 - \hat{M}_X(\theta_G)_i|M_i = 0)$ is then the value of interest

$$\mathbf{\Phi}_{\dot{X}}(x, 1 - \hat{M}_X(\Theta_G)_i|M_i = 0) = \frac{\sum_{i=1}^n (1 - \hat{M}_X(\Theta_G)_i) \cdot \mathbf{1}(G(\mathbf{X}|\Theta_G)_i \leq x) \cdot \mathbf{1}(M_i = 0)}{\sum_{i=1}^n (1 - \hat{M}_X(\Theta_G)_i)\mathbf{1}(M_i = 0)}$$

Note that $\hat{\mathbf{X}}$ is also a function of Θ_G: $\hat{\mathbf{X}} = \mathbf{M} \odot \mathbf{X} + (1 - \mathbf{M}) \odot G(\mathbf{X}|\Theta_G)$.

To ensure that $\mathcal{W}$ convergences when we estimate the proposed RWIGAN algorithm, we restrict the domain of the gradient of Θ_G, ∇_{Θ_G} with respect to $\mathcal{L}(\cdot)$. In the typical GANs, convergence is not guaranteed. For WGAN, the Kantorovich Rubinstein duality guarantees a unique global minima with respect to Θ_G. We seek to reinforce the notion by ensuring that every step of the descent algorithm is monotonic. We first derive the gradient of the quantile function $\mathbf{\Phi}$ for random variable X,

$$\nabla_G \mathbf{\Phi}(\dot{X}, 1 - \hat{m}_X, \Theta_G|M_i = 0) = \nabla_G \frac{\sum_{i=1}^n v_i(M)\mathbf{1}(G(\mathbf{X}|\theta_G)_i \leq x)}{\sum_{i=1}^n v_i(M)}$$

where for simplicity we write

$$v_i(M) = (1 - \hat{M}_{X,i})\mathbf{1}(M_i = 0),$$

This derivative exists as it is a sum and quotient of differentiable terms wrt Θ_G. By properties of the inverse derivative.

$$\nabla_G \mathbf{\Phi}^{-1}(\dot{X}, \hat{m}_X, \Theta_G|M_i = 0) = \frac{1}{\nabla_G \mathbf{\Phi}(\mathbf{\Phi}^{-1}(\dot{X}, 1 - \hat{m}_X, \Theta_G|M_i = 0)|M_i = 0)} \tag{11}$$

Note that the indicator function is not differentiable in the typical sense, but the discontinuity could be assumed away using the delta dirac function δ. We can replace this with a 'soft' continuous approximator of indicator:

$$\begin{aligned}
\mathbf{1}(G(\mathbf{X}, \Theta_G)_i \leq x) : &\approx \frac{\exp(C(x - G(\mathbf{X}|\Theta_G)_i))}{1 + \exp(C(x - G(\mathbf{X}, \Theta_G)_i))} \\
&:= S(\mathbf{X}|\Theta_G)
\end{aligned}$$

for some fixed constant C.

$$\nabla_G S(\mathbf{X}, \Theta_G) = -\frac{C \cdot \nabla_G G(\mathbf{X}, \Theta_G)_i \cdot \exp(C(x + G(\mathbf{X}, \Theta_G)_i)))}{\exp(C(G(\mathbf{X}, \Theta_G)_i) + \exp(C(x))^2}$$

so if $\nabla_G G(\mathbf{X}, \Theta_G)_i > 0$ then this derivative is positive, implying that the overall gradient for variable i is negative. In practice , to constrain the gradient to be positive, we can use *weight clipping* that was employed in the WGAN (Arjovsky et al. (2017)) to make sure the gradient is positive.

Figure 6 shows the convergence of G and D losses for a sample run. G appears to be smooth and convergence over 1000 runs , (Negative) D loss appears to increase at first while the generator loss has not yet attenuated, but then converges as generator converges. The convergence of the G loss, however, is the primary focus.

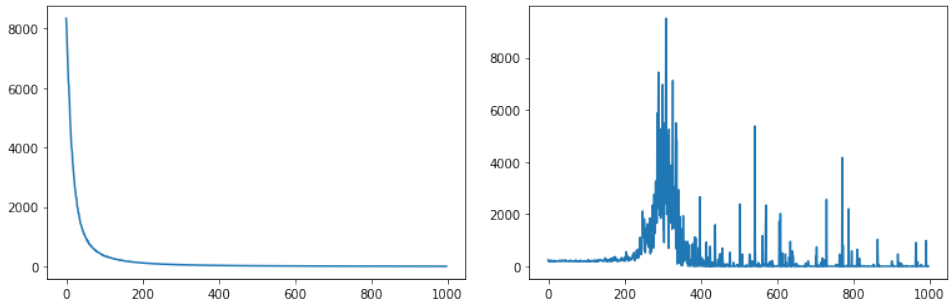

Figure 6: Convergence of G-loss and *negative* D loss in a sample run from a simulation

D REACH DETAILS

Approximately 20% of PCL variables from *raw* have missing entries that has been retroactively filled in or imputed. Moreover, exploratory analysis shows that the densities between the observed and missing (of which we know the ground-truth values of REACH *(raw)* appear inherently different (Figure 7 in Appendix). We also analyze *only* the *real missing values* of the *raw* dataset, whose ground-truth values are found in the smaller *cleaned* dataset.

There are two instantiations of the REACH data – the full dataset has 1776 subjects with 393 total variables. We call this REACH (a) (within the appendix). 62 of these variables are numeric and at least approximately continuous. Many variables are questionnaire responses on an ordinal (ranked) scale, but some responses are numerous enough as to be considered approximately 'quantitative'. We set the threshold of the number of (ranked ordinal) responses to be 9 for it to be considered quantitative for the purposes of imputation. Indeed, if the number of responses was only 5, for example, then the assumptions of heavy-tailedness in EHR data that would also apply to the Insight data (i.e. from Albers et al. (2018)) would not have much meaning.

The other is a smaller, more curated subset of the REACH (a). We call this REACH (b). Some of the data that is missing in (a) are filled in by subsequent additional observations in (b). While others imputed in (b) by expert imputations. There are instances where the data is missing in the raw, but observed in cleaned (imputed) version. These expert imputations are performed qualitatively; we assume that the mixture of these imputations with retrospective filled-in data are a surrogate for *ground truth*. We subset the data further to only *continuous* variables, which include the *EHT time to arrival*, many instantiations of *PCL* scores (that are not available in the smaller REACH (b) dataset), EDP, and *PHQ* scores. The resultant dataset has 62 variables with the full 1776 observations. The *ground truth* data has approximately 36% entries missing, but the induced-missing dataset as in method in Table 2.

D.1 MISSINGNESS ASSUMPTIONS OF REACH

The PCL Combo data in REACH is perhaps the most variegated, informative, and consequential sub-category of the REACH data. The PCL score is taken longitudinally in the larger dataset, with many different strata across different times. PCL data is also the only subset of data that has a

REACH Dataset Missingness Characteristics

Data	n	p	$\mathbb{P}$(missing)
REACH(a): Large (Raw)	1776	373	.30
REACH(b): Small (Cleaned)	764	99	.01
Processed REACH (a)	1776	62	.37
Processed REACH (a)(*induced missing*)	1776	62	.54
REACH (a)-PCL (*induced missing*)	1776	24	.64
REACH (b)-PCL (*real missing*)	764	4	.21

Table 3: Basic properties of the REACH dataset, and the resultant REACH datsets for analysis (bottom).The bottom 4 examples are used for actual imputations in Table 2.

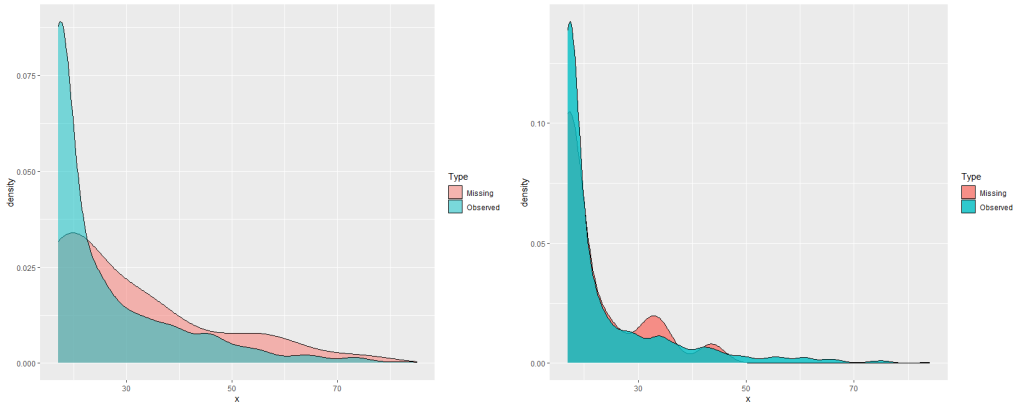

Figure 7: (left) Density plots for observed (i.e. same value in raw and cleaned datasets) and missing (i.e. retroactively filled-in in the "cleaned" dataset) PCL baseline and (right) PCL 12 months

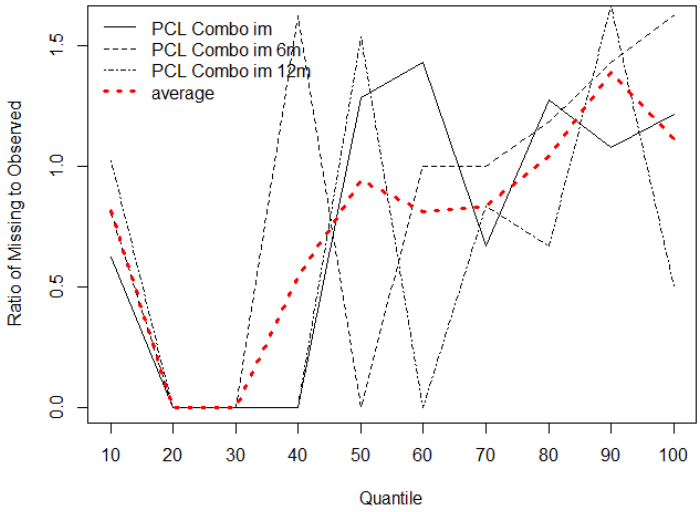

Figure 8: Ratios of the proportions of missing (i.e. Y_j^I) variables for PCL Combo scores at baseline, 6mo, 12mo, to those that are only observed (i.e. Y_j^O) , that fall within the decile levels as derived from the common dataset Y_j^C. The red line indicates an average of the PCL score ratios between missing (i.e. expert imputed) and observed . The red line shows a pattern of increasing missingness among the upper quantiles. This pattern implies that for PCL data (which in a sense undergirds the rest of the REACH data).

sizeable proportion of missing from the *raw* (i.e. REACH(a))that has been retroactively filled in or imputed. Approximately 20% of the PCL scores are missing from the raw data (REACH (a)), but there exists a substantial sample of data that fit this category. Moreover, upon inspection, exploratory visual analysis shows that the densities between the observed and missing (of which we know the ground-truth values of in REACH (b)) do look inherently different (Figure 7).

We model missingness in REACH based on the different forms of PCL data that is present in both datasets. As such, we posit that one way to model the missingness by using the patterns drawn from parallels the patterns evinced from the aggregate of the PCL scores. We estimate missingness rates by comparing the PCL variables from the *raw* REACH data where $n = 1776$ and the *cleaned* where $n = 764$. Indeed, prior studies place much emphasise on PCL variables Birk et al. (2019) . Let $\mathbf{Y}^I$ be the *expert imputed [PCL] data*, or retrospectively filled-in tru observations, and let $\mathbf{Y}^C$ be the observed data that is common to both *raw* and *cleaned*. We "bootstrap" the information in the discrepancy between the raw and cleaned.

For each PCL variable (baseline, 1, 6, and 12 months follow-up), we evaluate the relative ratios *observed* and *missing* (i.e. imputed) fixed quantile grids for the *common* dataset. We treat the quantiles $\tau_q(Y_j^C)$ of the common (i.e. identical entries between Reach (a) and (b) as the ground truth quantiles. We then calculate the proportion of the *observed* data that is only present in the larger dataset (REACH (a)) and the missing data that is only present in the smaller dataset)REACH (b)). We calculate what proportion of the above *missing* $\mathbf{P}_j^{\text{msg}}$ and *observed* $\mathbf{P}_j^{\text{obs}}$ points fall in the bins demarcated by $\tau_q(Y_j^C)$, for deciles $q = 0, 10, 20, ..., 100.$, and then take the ratio $\mathbf{R}_j := \mathbf{P}_j^{\text{msg}}/\mathbf{P}_j^{\text{obs}}$ of missing to observed proportions to derive a metric that measures the relative prevalence of value-ranges based on the missingness rates of the variable in question. Figure 8 shows the pattern of $\mathbf{R}_j$ for various PCL variables.

We posit that if the quantiles of $\mathbf{Y}^I$, $\tau_q(\mathbf{Y}^I)$ (at some fixed level q) are consistently higher than $\tau_q(\mathbf{Y}^C)$, then there is evidence of underestimation of the initial imputations, then the actual missingness rate may be higher. There is a sizeable difference between the quantiles between the missing entries in the PCL variables in Y_j^I. Indeed, most of the PCL variables that are imputed (or filled-in in some way) have higher concentrations in the upper quantiles. We posit that this

signals an *underestimate*. As such, we calculate the ratios between the missing and imputed, then fit a spline-like function through these ratios and fit a function that approximates this pattern. We then use this function to induce additional missingness in the data, which will be described in the following sub-section.

Let $\mathbf{Y}^I$ be the *expert imputed [PCL] data*, or retrospectively filled-in true observations, and let $\mathbf{Y}^C$ be the observed data that is common to both *raw* and *cleaned*. We "bootstrap" the information in the discrepancy between the raw and cleaned. Details on the design of inducing missingness is found in Appendix D.1. The missingness for a given X_j is modeled as:

$$\mathbb{P}(X_j \text{ is missing}) = \frac{\log(\kappa_j + 1)}{\max(\log(\kappa_j + 1))}$$

where $\kappa_j = \frac{(n_j/4 - \text{rank}(X_j))^2}{n_j}$. This function forms a check-shape in Figure 11. This pattern implies that at very low values, entries are more likely to be missing, but then after the quantile reaches 30 or 40%, the missingness begins to increase as the values get larger, with a probability of 1 as the5 values approach maximum.

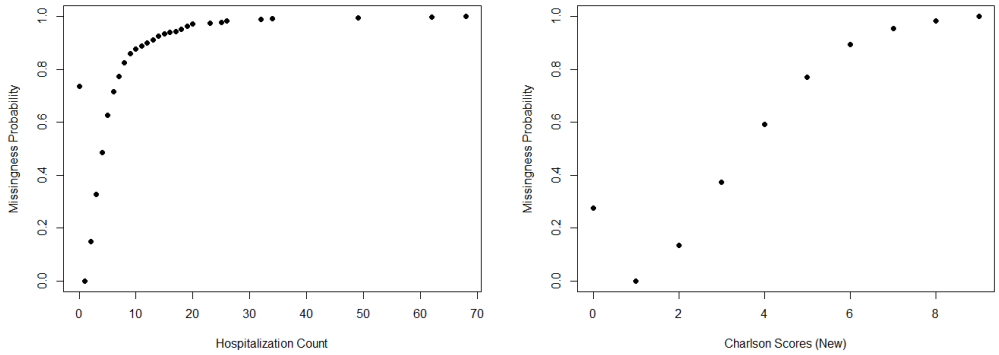

Figure 9: Induced missingness rates per decile for two sample variables *hospitalization count* and *Charlson Scores (new)* as the parallel the estimates of the PCL scores

D.2 REACH ANALYSIS DETAILS

We estimate missingness rates by comparing the PCL variables from the *raw* REACH data where $n = 1776$ and the *cleaned* where $n = 764$. Prior studies place much emphasis on PCL variables Birk et al. (2019). We model the missingness of PCL by using the difference in missingness between the *raw* and *cleaned* datasets. Approximately 20% of PCL variables from *raw* have missing entries that has been retroactively filled in or imputed. Moreover, exploratory analysis shows that the densities between the observed and missing (of which we know the ground-truth values of REACH (*raw)* appear inherently different (Figure 7 in appendix). We also analyze *only* the *real missing values* of the *raw* dataset, whose ground-truth values are found in the smaller *cleaned* dataset. Results show that RWIGAN is preferable to GAIN and MIDA consistently in terms of MSE, nMSE, as well as the (canonical) W_1 distance (defined in (1)). MICE does slightly better than RWIGAN in some cases. Although it is much slower. When subset to the smaller dataset, RWIGAN performs worse because the real rates of missingness are not as strongly defined as induced. Though we have reason to believe that the real missingnesss rates obscure hidden, and since the *cleaned* dataset is well curated, and the observed variables have hidden additional biases, it may be reasonable in real-world settings to assume that the modeled mechanism of MNAR is understating the true missingness.

D.3 INSIGHT DETAILS

For subjects with *positive* tests for COVID-19, we aggregated all *quantitative* (i.e. non-binary) laboratory tests which exceed 1000 tabulations (across the entire population) and took the median test value for subjects with repeated measurements. As such, the There are a total of 754 laboratory test codes that fit in this criteria. However, we remove those with too many missing (as to collapse most imputation algorithms) and set the minimum observed values *for variables* (i.e. columns) to

be 40, and the minimum observed values *for subjects* to be 400. This is to ensure that the data is not overly sparse so that imputation algorithms would actually run. The resultant dataset has 4764 total subjects (rows) and 162 variables (columns).

E Details for Simulation and Data Analysis

E.1 Evaluation Criteria

We use MSE and nMSE for Mean squared error (MSE) measures the distance between imputed value $\hat{X}_{ij}$ at row i and column j, and the ground-truth value X_{ij}. Each value has a mask M_{ij} that is equal to 1 if the data is observed and 0 if it is induced missing. n_j represents the number of *observed* points for variable j. This metric is serves as the mean distance of all the (true) missing values with their imputed values by the various methods.

$$\text{MSE}(\mathbf{X}_j) = \frac{\sum_{i=1}^{n_j}(\hat{X}_{ij} - X_{ij})^2(1 - M_{ij})}{n_j},$$

$$\text{nMSE}(\mathbf{X}_j) = \frac{1}{\sum_{k=1}^{n_j} X_{kj}/n_j} \sum_{i=1}^{n_j}(\hat{X}_{ij} - X_{ij})^2(1 - M_{ij}).$$

E.2 Details of Simulations

In the first set of IM simulations (IM-exp), each column is a vector indexed at j X_j is generatred from independent exp(.1) distributions. The missingness is:

$$\mathbb{P}(X_{ij} \text{ is missing}) = \frac{\log(X_j + 1)}{\max(\log(X_j + 1))}$$

In IM-u-a, each column X_j is a vector generated from unif(0,1). The missingness is 50 % if the value of the i-th entry of X_{ij} is over the 70 % quantile of the column.

In the first IMCI simulation (IMCI-a) , each X_j is drawn from a multivariate normal distribution with correlation $\rho = .25$ (in relation to the other X_j's) and each Y_j is composed of X_j with an unobserved independent standard uniform U_j: $Y_j \sim 1 + 2U_j + X_j + \exp(.5)$. The second set of IMCI simulations (IMCI-b) is comprised of independent exp(.01) variables X_j with Y_j which is dependent on X_j in the following way $Y_j \sim 1 + 2X_j + \exp(.1)$.

E.3 Additional Simulation Figures

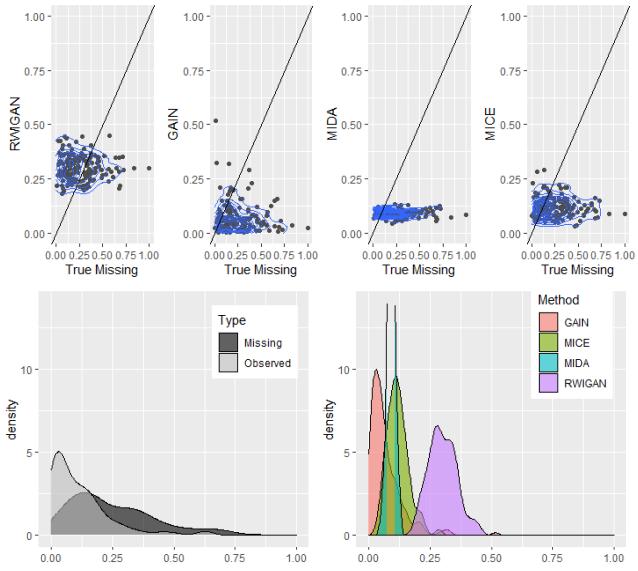

Figure 10: Another example of IM simulation imputation

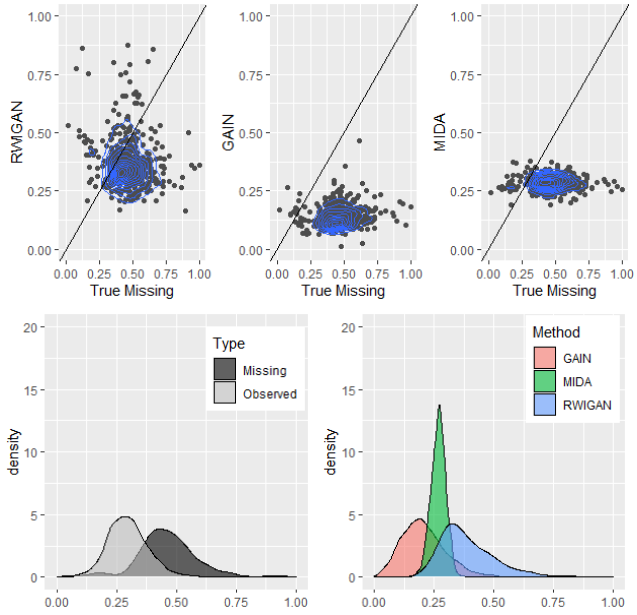

Figure 11: Another example of an INSIGHT variable *Cholesterol in HDL (mass volume) in blood*