# A neural network approach for large-scale imputation for data with informative missingness by minimizing re-calibrated Wasserstein distance

## 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 to enable efficient imputation in large-scale observational data in presence of informative missingness and covariate imbalance. We propose a new quantile re-weighting technique to ensure distributional equivariance under informative missingness and alternate estimation between a recalibrated Wasserstein distance objective with a logistic loss function for the estimated missingness. 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 tests from the INSIGHT and MIMIC datasets.

*Keywords: Wasserstein distance, neural network, multiple imputation, inverse probability weighting*

# 1 Introduction

Large-scale observational data, including *electronic health records* (EHR), have been increasingly used to inform clinical decisions and discover disease etiologies. However, these datasets often contain missing values due to various reasons, and the underlying factors that result in missingness convey meaningful information. For example, in an EHR system, 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., 2018); in psychological questionnaires, variables that measure the severity of depression or mania may be correlated to the likelihood of response. Missing data in EHR can strongly suggest potential underlying health conditions. Tan et al. (2023) studied informative missingness in EHR across 9 health systems and confirms the prevalence of non-random missingness. In the wake of the COVID-19 pandemic, the responsible use of big data for healthcare is an especially urgent and important issue. Such *informative* missing data have a considerable impact on the credibility and soundness of clinical evaluations if not adequately dealt with during analyses.

There are several ways to make missing data amenable for analysis. Among those, imputation is most commonly-used approach, and can better preserves the fidelity of the data in comparison to the complete-data-only analysis(Graham, 2009). The key for a successful imputation is a good estimation of conditional distribution of missing values given the existing observed variables.

There are several ways to make missing data amenable for analysis. Among those, imputation is most commonly-used approach, and can better preserves the fidelity of the data in comparison to the complete-data-only analysis(Graham, 2009). The key for a successful imputation is a good estimation of conditional distribution of missing values given the existing observed variables.

One commonly used imputation method is MICE, which 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 that outperforms other existing methods in various applications, but the biggest limitation is that computation time increases quadratically as with the number of variables (Wang et al., 2021). Alternatively, another class of imputation methods use deep learning (DL) and have been a recent trend due to its computational advantages. MIDA is one example based on denoising autoencoders (Gondara and Wang, 2017). GANs are another popular framework for imputation (Goodfellow et al., 2014); examples include GAIN and WGAIN (Friedjungová et al., 2020; Yoon et al., 2018; Li et al., 2019).

Yoon et al. (2018) and Yang et al. (2019) posit that (W)GAN-based approaches can be considered forms of *multiple imputation* because its generative aspect "models all features with missing values simultaneously". As such, they are popular modern tools for imputation. However, they are not without its limitations:Wang et al. (2021) demonstrates that MICE actually outperforms nearly all deep learning (DL) -based approaches for small to moderately sized datasets, in which MICE is still computationally feasible. (Gondara and Wang, 2017). The fundamental reason is – that most of these algorithms operate under the assumption that the missing data is completely random, and as a result, the observed and missing data have identical distributions. Broadly speaking, (W)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 model distribution $P_G$ that is as similar to 'real' distribution $P_{\text{data}}$. $D$ tells if what is generated by $G$ comes from $P_{\text{data}}$ rather than $P_G$. By iterating these competition and generation steps, (W)GAN eventually reaches a steady-state where $P_G$ approximates $P_{\text{data}}$ , which can then be used for the imputation of missing values. By design, (W)GANs generate missing values by mimicking observed data.

Inverse probability weighting (IPW) is a common strategy to handle the informative missingness (IM)(Wei et al., 2012; Yuan and Dong, 2019; Xie and Zhang, 2017). Luo et al. (2018) proposed a GAN-based imputation in time series by introducing a novel recurrent neural network cell called *gated recurrent unit* which uses a time-decay term to takes account for temporally correlated missingness, but require time-series data and do not account for MNAR in single-time observations. Mattei and Frellsen (2019) and Ipsen et al. (2020) uses importance weighted autoencoders to reweigh imputation values using a deep latent variational model (DVLM). Ipsen et al. (2020) specify the missingness model by modeling the missingness mask with the data. We In this paper, we propose a method under the MNAR assumption using a similar approach of alternating between generating the imputed values and modeling the missingness itself.

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 neural network (NN) based imputation method whose objective function is recalibrated by estimated missingness probabilities (2) application to high-quality real EHR data with informative missingness , together with simulations that approximate them. Our goal is to impute missing values in MNAR data. "Not-MIWAE" proposed by Ipsen et al. (2020) uses variational autoencoders to recover the distribution of the missing data conditional on its missingness rate. We approach the MNAR problem from a different angle and demonstrate empirically that our proposed method works better. We define the objective functions in Section 3. We name this new method "Generative Reweighted Wasserstein Adversarial Imputation Network", or *GRWAIN*. This method 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, Little and Rubin), by using a *recalibrated* Wasserstein distance that is reweighted by the estimated missing probability to achieve better fidelity and minimize bias in reconstruction. We describe this in Section 3. However, unlike the existing GAN-based imputation approach of Yoon et al. (2018), we prove the monotonicity and convergence of the global Wasserstein criterion, as well as identifiability and consistency of the empirical CDFs and their associated missingness weights. In Section 4.1, we demonstrate that our proposed method outperforms other existing methods in (ordinary) $W_1$ distance minimization on real data, following evaluation of simulations.

## 2 Methodology

Suppose $\mathbf{X} = (X_1, ..., X_p)$ is a $p$ dimensional vector of variables of interest with a joint density $f(\mathbf{x})$, and $\mathbf{M} = (M_1, ..., M_p)$ is its coupling vector of indicators of its missingness. In other words, $X_j$ is observed when $M_j = 1$, and it is considered missing when $M_j = 0$. Within the field of machine learning, $M$ is often referred to as the mask, a mechanism that conceals the actual values. Ideally, all elements of $\mathbf{X}$ would be observed. However, within the context of observational Electronic Health Record (EHR) data, only a subset of these elements is observable. Consequently, the observed data can be represented as $\mathbf{X}^{obs} = \mathbf{X}[\mathbf{M} = 1]$, while the missing (or unobserved) data can be encapsulated as $\mathbf{X}^{miss} = \mathbf{X}[\mathbf{M} = 0]$. An *informative* missingness scenario arises when the probability of missingness depends on the actual value of the variable, i.e., $\mathbb{P}(M = 1|X) = \pi(X)$. Under these circumstances, the distribution of the observed data deviates from that of the missing data. This creates a challenge for machine learning algorithms that assume the equivalence of these distributions, such as (W)GAIN, potentially introducing a bias into their performance and outcomes.

## 2.1 Distributional Equivalence under Informative Missingness and its Identifiability

Suppose $f(x)$ is the marginal distribution of $X$, $f^{obs}(x) = f(x \mid M = 1)$ signifies the conditional density of $X$ when observed, and $f^{miss}(x)$ denotes the density when $X$ is missing. Leveraging Bayes' theorem, we have:

$$\frac{f^{obs}(X)}{\pi(X)} \propto f(X), \text{ and } \frac{f^{miss}(X)}{1-\pi(X)} \propto f(X). \tag{1}$$

These relations suggest that a form of distributional equivalence between $f^{obs}(X)$ and $f^{miss}(X)$ can be achieved by *redistributing* their probability mass according to the missing probability $\pi(x)$. To show this with a simple example, we generate a random sample $(x_i, M_i)$ with informative missingness, where $x_i$ follows a $U(0, 1)$, and $M_i$ is a binomial distribution with prob$(M_i = 1) = \log(x_i + 1)$. In Figure 1, we present the QQ plot of original empirical distributions of observed and missing $x_i$'s (in black dots), and that of re-weight ones using the true weights(in red dots). As shown, the original distributions of observed and missing $x_i$'s are different, while the probability mass redistribution can lead to distributional equivalence. This re-weighted distributional equivalence allows for the adaptation of generative neural network (NN) algorithms, and is the cornerstone of the proposed algorithm

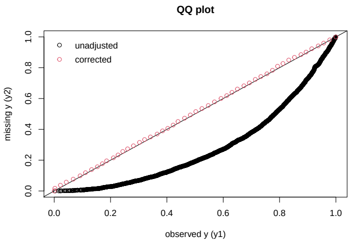


Figure 1: QQ plots

Prior to detailing the construction of generative NN algorithms based on this reweighted distributional equivalence, it is critical to address an identifiability issue, namely, whether the two unknown functions $f^{msg}(x)$ and $\pi(x)$ can be identified and estimated from the observed data $(X_i, M_i)$.

The marginal distribution $f(x)$ can be interpreted as a mixture of $f^{obs}(x)$ and $f^{miss}(x)$ such that

$$f(x) = \pi_0 f^{obs}(X) + (1 - \pi_0) f^{miss}(X) \tag{2}$$

where $\pi_0 = \int \pi(x)f(x)dx$ is the marginal observedness probability. In this equation, both $\pi_0$ and $f^{obs}(x)$ can be estimated consistently and nonparametrically from the data. Thus, the identifiability of $f^{msg}(x)$ and $\pi(x)$ is equivalent to the identifiability of $f(x)$ and $\pi(x)$. The likelihood of the observed data, denoted as $(x_i, M_i)$, we can gain valuable insights into the direct information available from the data. This likelihood can be expressed as

$$[f(x_i)\pi(x_i)]^{M_i}[1 - \int f(x)\pi(x)dx]^{1-M_i}, \tag{3}$$

which is a function of $f(x)\pi(x)$.

Equation (3) suggests that the product of $f(x)\pi(x)$ is identifiable and can be estimated nonparametrically from the observed data, but not these two functions separately. Some parametric assumptions for either $f(x)$ or $\pi(x)$ are necessary to ensure the identifiability. That is, we assume that either $f(x)$ or $\pi(x)$ is a parametric function and satisfies the following conditions.

**Condition 1:** The missing rates are bounded away from 0 and 1, i.e. $0 < \pi(x) < 1$ for any $x$.

**Condition 2:** $f(x) = f(x, \boldsymbol{\theta})$ and if $f(x, \boldsymbol{\theta})\pi(x) \neq f(x, \boldsymbol{\theta}^*)\pi(x)$ , then $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$

**Condition 2':** $\pi(x) = \pi(x, \boldsymbol{\theta})$ and if $f(x)\pi(x, \boldsymbol{\theta}) \neq f(x)\pi(x, \boldsymbol{\theta}^*)$ , then $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$

$\boldsymbol{\theta}^0(\cdot)$ represents a general operator for the ground truth parameters, which could take the form of coefficients, intercepts, or empirical quantile functions. Condition 1 is necessary to ensure that $\pi(x)$ is invertible. This condition also implies that $f^{obs}(x)$ and $f^{miss}(x)$ share the same support, despite potentially differing in shape. Condition 2 or 2' guarantee that the two density functions $f^{miss}(x)$ and $\pi(x)$ are identifiable. Although we assume that one of $f(x)$ and $\pi(x)$ is a parametric function, there is no limit on the dimension of $\boldsymbol{\theta}$ to ensure its practical flexibility. Later, in Section 3.1, we specify the (general) parameters $\boldsymbol{\theta}$ to be $\Theta_G, \Theta_D$, the parameters for fully-connected neural networks.

## 2.2 Generalized empirical distribution and quantile quantile functions and their Wasserstein Distances

Let $X = (X_1, ..., X_n)$ be a vector of $n$ observations, and $\boldsymbol{\omega}$ be a vector of probability mass that satisfies $0 \leq \omega_i \leq 1 \ \forall i$ and $\sum_i \omega_i = 1$. We define a class of *generalized* empirical distribution by

$$\tilde{F}_X(x; \boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i I\{X_i \leq x\}.$$

Under this definition, the traditional empirical distribution is a special case, with all the data points in $X$ receiving equal probability mass $1/n$. Let $\{x_{(1)}, x_{(2)}, ..., x_{(n)}\}$ be the order statistics of $X$, and $(\omega_{(1)}, ..., \omega_{(i)}, ..., \omega_{(n_1)})$ be their corresponding re-assigned probability mass. The *generalized* empirical quantile function can be written as

$$\tilde{Q}_X(\tau; \boldsymbol{\omega}) = \inf\{x_{(k)} : \sum_{i=1}^{k} \omega_{(i)} \geq \tau; \ k = 1, ..., n\}.$$

The Wasserstein metric is a robust and flexible metric that quantifies the 'distance' between two probability distributions. Given two probability measures $\mu$ and $\nu$ on a metric space $\mathcal{X}$, the $p$-Wasserstein distance $W_p(\mu, \nu)$ is defined as:

$$W_p(\mu, \nu) = \left( \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p d\gamma(x, y) \right)^{1/p}, \tag{4}$$

where $\Gamma(\mu, \nu)$ is the set of all joint distributions $\gamma(x, y)$ on $\mathcal{X} \times \mathcal{X}$ with marginals $\mu$ and $\nu$ and $d(x, y)$ is a distance function on $\mathcal{X}$. The metric calculates the minimum 'work' needed to transform one distribution into the other, where 'work' is defined as the product of the amount of mass moved (i.e. $d\gamma(x, y)$) and the distance it is moved (i.e. $d(x, y)$). This way, it provides more sensitive, intuitive and meaningful measure of distributional disparity than other metrics like the Kullback-Leibler divergence and $L_2$ distance. Furthermore, the distances possess useful mathematical properties, such as continuity and differentiability, which are advantageous in optimization, especially in high-dimensional machine learning problems. In the context of Wasserstein GAN, $W_1$ between two distributions $X$ and $Y$ is typically defined as its Kantorovich dual as the supremum across the difference of expected values of observations from these two distributions, after being mapped by any 1-Lipschitz function (Arjovsky et al. (2017)).

When the distributions $\mu$ and $\nu$ are univariate, 1-Wasserstein distance is

$$W_1(\mu, \nu) = \int_0^1 |\mu^{-1}(\tau) - \nu^{-1}(\tau)| d\tau,$$

where $\mu^{-1}(\tau)$ and $\nu^{-1}(\tau)$ are the quantile functions of $\mu$ and $\nu$. $W_1$ can be viewed as the minimum cost of transporting mass in order to transform the probability distribution $\mu$ into the distribution $\nu$ at quantile level $\tau$ between 0 and 1. We define generalized 1-Wasserstein distance to incorporate weights. Let $\mathbf{X}$ and $\mathbf{Y}$ are two random samples, whose coupling probability mass vectors are $\boldsymbol{\omega}_x$ and $\boldsymbol{\omega}_y$ respectively. Their generalized empirical quantile functions are $Q_{\mathbf{X}}(\tau;\boldsymbol{\omega}_x)$ and $Q_{\mathbf{Y}}(\tau;\boldsymbol{\omega}_y)$, and the generalized 1-Wasserstein distance is

$$\mathcal{W}_1(\mathbf{X},\mathbf{Y},\boldsymbol{\omega}_x,\boldsymbol{\omega}_y) = \int_0^1 |Q_{\mathbf{X}}(\tau;\boldsymbol{\omega}_x) - Q_{\mathbf{Y}}(\tau;\boldsymbol{\omega}_y)|d\tau. \tag{5}$$

## 2.3 Objective function

The objective function of the proposed algorithm is to find optimal $\mathbf{F}^{(miss)}(\mathbf{X})$ (i.e., the distribution of missing data) and $\boldsymbol{\pi}(\mathbf{X})$ (i.e., the missing probability) that minimize the distance between the *re-calibrated* $\mathbb{P}_{obs}$ and $\mathbb{P}_{mis}$. To do so, we propose a Wasserstein Quantile-Quantile Distance. Let $\mathbf{F}_{\mathbf{X}}$ and $\mathbf{F}_{\mathbf{Y}}$ be two multivariate distribution functions representing $n \times p$ dimensional multivariate distributions $\mathbf{X}$ and $\mathbf{Y}$. We measure their difference by aggregating over marginal differences as

$$\mathcal{W}_1(\mathbf{F}_{\mathbf{X}},\mathbf{F}_{\mathbf{Y}},\omega_{\mathbf{X}},\omega_{\mathbf{Y}}) = \sum_{j=1}^{p} W_1(Q_{\mathbf{X},j}(\omega_{\mathbf{X}}),Q_{\mathbf{Y},j}(\omega_{\mathbf{Y}})) = \sum_{j=1}^{p} \int_0^1 \left|Q_{\mathbf{X},j}(\tau,\omega_{\mathbf{X},j}) - Q_{\mathbf{Y},j}(\tau,\omega_{\mathbf{Y},j})\right|d\tau, \tag{6}$$

where $Q_{\mathbf{X},j}$ and $Q_{\mathbf{Y},j}$ are the $j$th marginal quantile functions of $\mathbf{F}_{\mathbf{X}}$ and $\mathbf{F}_{\mathbf{Y}}$. In comparison to the more common measures, such as Kullback–Leibler (KL) divergence, quantile-quantile distance does not require the distributions to have common support, and hence, is more suitable and robust for heterogeneous data like EHR. In addition, the quantile-quantile difference can be viewed as a special case of Wasserstein distance, which has a natural connection with the min-max GAN optimization to ensure the computational and theoretical validity Villani (2008). The global objective is to minimize the above criterion $\mathcal{W}$.

In the following analogy to *observed* and *missing* segments of data, $\mathbf{F}_{\mathbf{X}},\mathbf{F}_{\mathbf{Y}}$ are representative of $\mathbf{F}^{(obs)},\mathbf{F}^{(miss)}(\Theta_G)$. Let $\mathbf{F}^{(impu)}(\Theta_G)$ represent the approximation of $\mathbf{F}^{(miss)}$ the parameters $\Theta_G$ of the *generator*:

$$\mathbf{F}^{(impu)}(\Theta_G) = \hat{\mathbf{F}}^{(miss)}$$

we control $\mathbf{F}^{(impu)}$ with the parameters $\Theta_G$ and $\Theta_D$, which characterize the neural networks $D$ and $G$. In the proposed framework, we set $\omega_{\mathbf{X}}$ as $\hat{\pi}(\hat{\mathbf{X}})$, the estimated observedness, and $\omega_{\mathbf{Y}}$ as the estimated missingness

$$\omega_{\mathbf{X}} := \hat{\pi}(\hat{\mathbf{X}}) \tag{7}$$

$$\omega_{\mathbf{Y}} := 1 - \hat{\pi}(\hat{\mathbf{X}}) \tag{8}$$

$\hat{\pi}(\hat{\mathbf{X}})$ learned by the Discriminator $D$ to approximate the missing probability $\boldsymbol{\pi}(\mathbf{X})$, while using the Generator $G$ to learn the distribution of missing data $\mathbf{F}^{(mis)}$ and generate imputations accordingly.

To estimate the NN parameters, we iteratively update $\Theta_D$ and $\Theta_G$ to minimize the following overall objective function

$$(\widehat{\Theta}_G,\widehat{\Theta}_D) = \min_{\Theta_G,\Theta_D} \mathcal{W}_1(\hat{\mathbf{F}}^{(obs)},\hat{\mathbf{F}}^{(impu)}(\Theta_G),\hat{\pi}(\Theta_D),1-\hat{\pi}(\Theta_D)) \tag{9}$$

where $\widetilde{\mathbb{P}}_{(obs)}(\Theta_D)$ is the distribution of the observed data re-calibrated by the $\Theta_D$-determined missing probability, and $\widetilde{\mathbb{P}}_G(\Theta_G,\Theta_D)$ is the distribution of imputed data $G(\mathbf{X},\Theta_G)$ re-calibrated by the $\Theta_D$-determined missing probability, $\mathcal{W}_1$ is the Wasserstein quantile-quantile distance defined earlier. Weights $\omega_x,\omega_y$ are implicitly included in the definition of $\mathcal{W}_1(\cdot)$ from (5) through $\Theta_D,\Theta_G$. The practical implementation is outlined in Section 3.

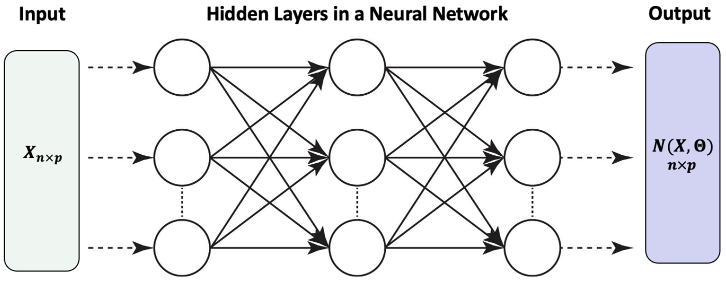


Figure 2: Diagram of neural network, which could represent either generator $G$ or discriminator $D$

# 3 Neural Network Model and Algorithm

Neural networks, as described in Section 1, have recently become a popular tool to impute missing data in high dimensions. Yoon et al. (2018) proposed a method of using GANs to impute data missing at random. The method uses the generator to create imputations, but uses the discriminator to predict the missingness rate as to rate the quality of the generated imputations. These methods comprise a framework for Deep-Learning based imputation methods: MISGAN also use the $D$ of the GAN to the mask probability. These are not quite GANs in that the discriminator does not judge the quality of generation itself, but rather the estimates of the missingness probabilities (Goodfellow et al. (2014)). These approaches have become popular in recent years because of their computational advantages for large scale data, but for most existing data forms, traditional techniques such as MICE (Wang et al. (2021)) are still preferable because of their accuracy. Noting the weaknesses of the NN based methods in accounting for MNAR, we augment these shortcomings by augmenting the Wasserstein minimization with recalibration of quantiles by the estimated missingness probabilities. We had introduced the reweighted Wasserstein distance for a general imputation setting. Now we specify that setting to the GAN setting off two alternating neural networks $G$ and $D$. $G$ is the generator with associated parameters $\Theta_G$ and $D$ is the discriminator with $\Theta_D$. Let

$$\hat{\mathbf{X}} := \mathbf{X}^{obs} \odot \mathbf{M} + G_{\Theta_G}(\mathbf{X}^{obs}, \mathbf{Z}) \odot (1 - \mathbf{M}) \tag{10}$$

be the output from the generator $G$. The criterion for GRWAIN is comprised of two parts: minimizing the $D$ loss and minimizing the $G$ loss. These steps are alternated and a schematic diagram for the implementation is shown in Figure 3. The likelihood for $D$, following the log form of Equation (3), is

$$\max_{\Theta_D} \mathcal{L}_D(\Theta_D|\Theta_G) = \max_{\Theta_D} \mathbb{E}_{\hat{\mathbf{X}},\mathbf{M}} \left[ \sum_{i=1}^{n} \left\{ \mathbf{M}_i \log(\hat{\pi}_i(\hat{\mathbf{X}}, \Theta_D)) + (1 - \mathbf{M}_i) \log(1 - \hat{\pi}_i(\hat{\mathbf{X}}, \Theta_D)) \right\} \middle| \Theta_G \right].$$

Alternating with the maximization of $\mathcal{L}_D$, we minimize the reweighted Wasserstein distance $\mathcal{W}$ between the observed data and imputed values.

$$\min_{\Theta_G} \mathcal{L}_G(\Theta_G|\Theta_D) = \min_{\Theta_G} \mathcal{W}_1 \left\{ \widetilde{\mathbf{F}}_{(obs)}, \widetilde{\mathbf{F}}_{(impu)}(\Theta_G) \middle| \Theta_D \right\} \tag{11}$$

Note that the above objective (6) depends on $\Theta_G$ in the following way:

$$\mathcal{W}_1 \left\{ \widetilde{\mathbf{F}}_{(obs)}, \widetilde{\mathbf{F}}_{(impu)}(\Theta_G) \middle| \Theta_D \right\} = \int_0^1 |\widetilde{Q}_{\mathbf{X}^{obs}}(\tau; \hat{\pi}(\hat{\mathbf{X}})) - \widetilde{Q}_{\dot{\mathbf{X}}(\Theta_G)}(\tau; 1 - \hat{\pi}(\hat{\mathbf{X}}))| d\tau$$

where $\mathbf{X}^{impu}$ is the output from the generator.

$$\dot{\mathbf{X}}(\Theta_G) = G_{\Theta_G}(\mathbf{X}^{obs}, \mathbf{Z})$$

where each *hidden* layer at $h_l$ at layer $l$. $Q_{\mathbf{X}^{impu}}(\tau; 1 - \hat{\pi}(\hat{\mathbf{X}}))$ is its quantile function at level $\tau$ with *fixed* weights $1 - \hat{\pi}(\hat{\mathbf{X}})$ (the estimated rate of missingness). The $D$-loss is constructed from the identification rate of $\hat{\pi}$ for *observed* or *missing*. It increases when $\hat{\pi}_i(\hat{X})$ is greatest when the data are observed, and decreases $\hat{\pi}_i(\hat{X})$ is smallest when the data are missing. The bigger the difference between $\mathbf{X}^{obs}$ and $\mathbf{X}^{impu}$ , the better $D$ is able to discriminate, as $G$ in turn becomes better at generating samples. A flowchart of the relationships between these variables is found in Figure 3.

## 3.1 Neural Network Architecture

Now we detail the algorithm in the neural network case. Let $\mathbf{X}$ be $n \times p$ data matrix, and $\mathbf{M} = \{\delta_{i,j}\}_{n \times p}$ be the mask matrix indicating the locations of missing data. The proposed algorithm uses two fully-connected neural networks (Figure 2) to iteratively train the distribution of missing data $\mathbf{F}^{(mis)}$ and the missing probability function $\boldsymbol{\pi}(\cdot)$ until re-calibrated distributional equivalence in the data is established. Specifically, we have the imputation generator $G$, written as:

$$G(\mathbf{X}) = G_{\Theta_G}(\mathbf{X}) = \sigma(\Theta_G^{W,L} H_G^{L-1}(\mathbf{X}) + \Theta_G^{b,L}),$$

whose input is the matrix $\mathbf{X}$, and output is the matrix of imputed values. $G$ is characterized by parameters $\Theta_G$ $\sigma(\cdot)$ is the sigmoid activation function at the topmost layer and each $H_l$ at layer $l = 1, .., L$ is

$$H_G^l(X) = \begin{cases} \text{relu}(\Theta_G^{W,l} \cdot H_G^{l-1}(X) + \Theta_G^{b,l}) & \text{if } l > 1 \\ \text{relu}(\Theta_G^{W,1} \cdot X + \Theta_G^{b,1}) & \text{if } l = 1 \end{cases} \tag{12}$$

In practice, we set the number of layers $L$ to be 3 and $\sigma(\cdot)$ is the *sigmoid* activation function when the data is normalized between 0 and 1, and *relu* otherwise. Discriminator $D_{\Theta_D}$, has input $\widehat{\mathbf{X}}$, a matrix of a mixture of imputed and observed values. The output is the matrix of missing probabilities. The activation function of the outermost layer of neural network $D$ is also *sigmoid*, as to best model probabilities. The hidden layers that comprise the internal structure of $D$ is the same in form as that of $G$ as written in (12). $G$ and $D$ both map input of $p \times n$ to $p \times n$ output, with hidden layers of weights (e.g. coefficients) with dimension $p \times p$ and biases (intercepts with . As such, both neural networks have $p^2$ weights (coefficients) and $p$ biases (intercepts) for each layer. $\hat{\pi}(\hat{X})$ is the direct output of $D_{\Theta_D}$ , but for $G$, it is the empirical distribution of the output, e.g. a transformation.

The method is *adversarial* in that the maximization of the logistic function for estimation of $\hat{\pi}$ is alternated with minimization of the distance function $\mathcal{W}$. However, like that of Yoon et al. (2018), it is not a "canonical" WGAN in that it does not train on a minimax objective by exploiting the Kantorovich duality (Villani (2008)). The adversariality comes from the maximization of the (concave) logistic function, which is equivalent to alternating minimization of two convex functions. As such, we do not call this method a typical (W)GAN, but instead "Generative Reweighted Wasserstein Adversarial Imputation Network": *GRWAIN*.

In the next lemma we posit that the optimal discriminator $D^*$ with respect to the optimal $\Theta_D^*$ , given a $\Theta_G$ is equivalent to the observedness probability $\hat{\pi}(\hat{X})$. A similar lemma exists in Yoon et al. (2018). This fact is crucial in proving the monotonicity of the global objective function of $\mathcal{W}$ minimization.

**Lemma 1.** *Let $P(\cdot, \mathbf{M})$ be the joint density between observation $\mathbf{X}^{obs}$ or imputation $\dot{\mathbf{X}}$ and $\mathbf{M}$. Then for fixed generator function $G$ with parameters $\Theta_G$, the optimal discriminator $D^*(\hat{X})$ with $\Theta_D^{*;(t)}$ that maximizes $\mathcal{L}_D$ given $\Theta_G^{(t)}$ at iteration $t$ is given by*

$$\begin{aligned} D^*_{\Theta_D}(\hat{\mathbf{X}}|\Theta_G) &= \frac{P(\hat{\mathbf{X}}, \mathbf{M}=1|\Theta_G)}{P(\hat{\mathbf{X}}, \mathbf{M}=1|\Theta_G) + P(\hat{\mathbf{X}}, \mathbf{M}=0|\Theta_G)} \\ &:= \hat{\pi}(\hat{\mathbf{X}}|\Theta_G) \end{aligned} \tag{13}$$

*Proof.* Appendix A. □

## 3.2 GRWAIN Algorithm

In the proposed adversarial framework, we use Discriminator $D$ to generate the weights $\hat{\pi}_i$ and use Generator $G$ to generate imputations. We iteratively update $\Theta_D$ and $\Theta_G$ to achieve a global optimization in re-weighted distributional equivalence. After normalizing the data by standardizing every column such that its maximum value is 1 and minimum is 0, the algorithm is initialized with *Xavier*-initiated values (initialization commonly used for neural networks, such that variances of the activations are same across every layer) for all of the coefficient parameters (e.g. *weight*), and zero for the intercepts (i.e. *bias.* Then stochastic gradient descent (SGD) is used to alternatively minimize $\mathcal{L}_G$ and maximize $\mathcal{L}_D$. In each $D$ step, we maximize the logistic function.

For simplicity in this section, as is done in Ipsen et al. (2020), we describe operations without loss of generality on vector $X$ that could represent an arbitrary vector of matrix $\mathbf{X}$. We describe each step $t$ of SGD in the following list of steps:

Step 1: Indices $B$ are randomly batched and subsampled from the full dataset. For the rest of the steps, assume that all vectors from the data $1, ..., n$ are subsampled from $B$

Step 2 $\mathcal{L}_D$-maximization: Using fixed $\Theta_G^{(t)}$, we update $\Theta_D^{(t)}$ Feed $X$ into the neural network $G$ to obtain imputed $\dot{X}^{(t)}$ at iteration $t$:

$$\dot{X}^{(t)}(\Theta_G) = G_{\Theta_G^{(t)}}(X^{obs}, Z|\Theta_D^{(t-1)})$$

using generator $G$ with given $\Theta_G^{(t)}$. Using fixed generator parameters $\dot{X}^{(t)}(\Theta_G^{(t)})$, calculate gradient $\nabla_{\Theta_D}\mathcal{L}_D(\hat{X}, M, \hat{\pi}(\hat{X}, \Theta_D))$ to update the parameters $\Theta_D^{(t+1)}$ in order to maximize $\mathcal{L}_D$ using SGD. Then using $\dot{X}^{(t)}$ , generate the mixed imputation at $\hat{X}^{(t)}$ at time $t$

$$\hat{X}^{(t)} = X \odot M + (1 - M) \odot \dot{X}^{(t)}(\Theta_G).$$

Then using this imputed value we can calculate the $\hat{\pi}(\hat{X})^{(t)}$. Note that from Lemma 1 that the optimal discriminator exists in a closed form:

$$\hat{\pi}(\hat{X})^{(t)} = D^*_{\Theta_D^{(t)}}(\hat{X}^{(t)}|\Theta_G^{(t-1)}).$$

Step 3: Noise $Z$ are sampled from a uniform (0,1) distribution (as the data are standardized between 0 and 1). For each index $i$, If the data are observed, then the observed data $X_i^{\text{obs}}$ are used. If it is missing, then it is filled in with noise $Z_i$ at index $i$. As such vector $X$ is a mixture of observed data and noise:

$$X := X^{\text{obs}} \odot M + (1 - M) \odot Z \tag{14}$$

Step 4 $\mathcal{L}_G$-minimization: Using fixed Using fixed $\Theta_D^{(t)}$, we update $\Theta_G^{(t)}$: using by taking the gradient of the inverse quantile recalibrated form of $\mathcal{W}_1$

$$\begin{aligned}\min_{\Theta_G} \mathcal{W}_1(\Theta_G) &= \min_{\Theta_G} \mathcal{W}_1\Big(\Theta_G), \hat{\pi}(\hat{X})^{(t)}, 1-\hat{\pi}(\hat{X})^{(t)}\Big) \\ &= \min_{\Theta_G} \int_0^1 |\widetilde{Q}_{X^{obs}}(\tau; \hat{\pi}(\hat{X})^{(t)}) - \widetilde{Q}_{\dot{X}^{(t)}(\Theta_G)}(\tau; 1-\hat{\pi}(\hat{X})^{(t)}|d\tau\end{aligned}$$

Steps 1 - 4 are repeated until the respective likelihoods $\mathcal{L}_G$ and $\mathcal{L}_D$ attenuate, then the final imputation is obtained by averaging the last 20 imputations to derive a stable estimate by means of multiple imputation.

The re-weighted empirical quantile functions can be written as

$$\widetilde{Q}_{X^{obs}}(\tau, \pi(\hat{X})) = \inf\left\{X_{(k)}^{obs} : \left\{\sum_{i=1}^{k} \frac{1/\pi(\hat{X})}{\sum_{k=1}^{n_1} 1/\pi(\hat{X}_k)}\right\}_{(i)} \geq \tau;\ k = 1, ..., n_1\right\}, \text{ and} \tag{15}$$

$$\widetilde{Q}_{\dot{X}}(\tau, 1-\pi(\hat{X})) = \inf\left\{\dot{X}_{(k)}(\Theta_G) : \left\{\sum_{j=1}^{k} \frac{1/(1-\pi(\hat{X})}{\sum_{k=1}^{n_0} 1/(1-\pi(\hat{X}_k))}\right\}_{(j)} \geq \tau;\ k = 1, ..., n_0\right\} \tag{16}$$

we use a recalibrated Wasserstein Distance to evaluate quantile equivalence in the proposed imputation algorithm. We follow the approach of prior work such as Yuan and Dong (2019); Wei et al. (2012), where empirical likelihoods are reweighted by missingness probabilities. This technique is known as *inverse probability 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)). This procedure is described in detail in the following algorithm:

1. Partition $X$ into observed $X^{\text{obs}}$ and missing data imputations from output of $G$. Let the imputation $\dot{X} = G(X^{\text{obs}}, Z)$ represent the output of the generator $G$. Write $\hat{X}$ as the mixture of observed and imputed as in (10). Estimate missingness probabilities as: $\hat{\pi}_i(\hat{X}) = D(\hat{X}|\Theta_D)_i$. For any of the variables $\pi_i(\hat{X})$, $M_i$, $X_i$, $V_i$, $n$,$q$, let $^*$ denote either their observed or missing partition.

2. Define each $V_i^*$ (with associated length $n^*$ for either observed or missing):

$$V_i^* := \frac{1/\hat{\pi}_i(\hat{X})}{\sum_{i=1}^{n1} 1/\hat{\pi}_i(\hat{X})} \text{ or } \frac{1/(1-\hat{\pi}_i(\hat{X}))}{\sum_{i=1}^{n^1} 1/(1-\pi_i(\hat{X}))}$$

3. Now we take the observed and imputed (if missing) values of variable $X$. *Rank* each $X^*$ (in the observed or missing regime) and then define each $V_{(l)}^*$ as $V_i^*$ that corresponds with the index of $X_{(l)}^*$, the $l$-the ranked $X^*$. Define each quantile *height* $q_k^*$ ($*$ for either observed or missing) to be:

$$q_1^* = \frac{V_{(1)}^*}{\sum_{l=1}^{n} V_{(l)}^*}, ..., q_h^* = \frac{\sum_{k=1}^{h} V_{(k)}^*}{\sum_{l=1}^{n^*} V_{(l)}^*}, ..., q_{n^*-1}^* = \frac{\sum_{k=1}^{n^*-1} V_{(k)}^*}{\sum_{l=1}^{n^*} V_{(l)}^*}, 1$$

4. For a range of $\tau_k$, where $k = 1, ..., K$, the quantile functions $\widetilde{Q}_n^{obs}(\tau, \pi(\hat{X}))$ and $\widetilde{Q}_n^{msg}(\tau, 1-\pi(\hat{X}))$ from (15) are calculated. Choose from evenly gridded quantiles $\boldsymbol{\tau} = \{\tau_1, ..., \tau_K\}$, then linearly interpolate the $K$ points, obtain their quantile difference to otain the Wasserstein distance

▷ omit this and combine with previous algorith, specifically: how to combine $B$ generator and new patients?

The diagram outlined Figure 3 shows a flow of the procedure.

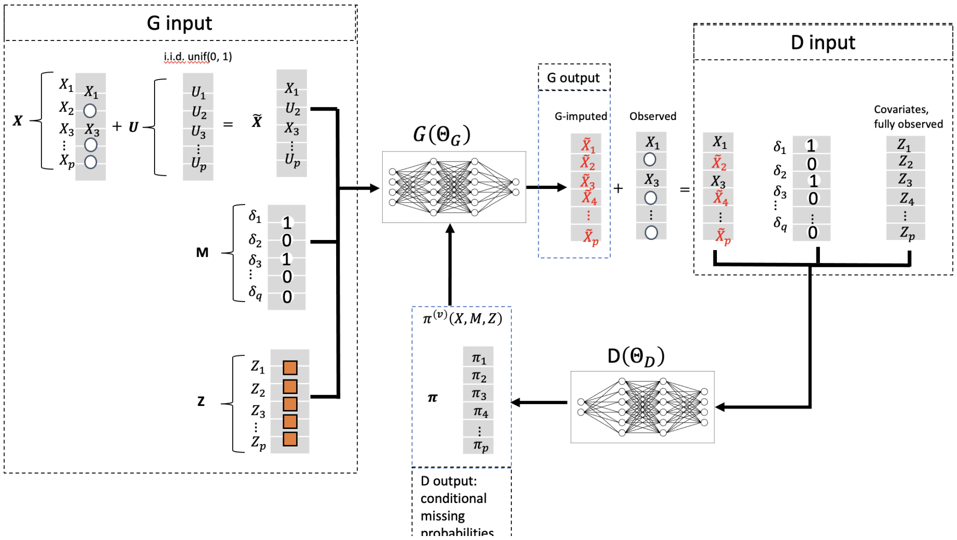

G input
i.i.d. unif(0, 1)
$X$
$X_1$ $X_2$ $X_3$ ⋮ $X_p$
$X_1$ ○ $X_3$ ○ ○
$+ U$
$U_1$ $U_2$ $U_3$ ⋮ $U_p$
$= \tilde{X}$
$X_1$ $U_2$ $X_3$ ⋮ $U_p$
M
$\delta_1$ $\delta_2$ $\delta_3$ ⋮ $\delta_q$
1 0 1 0 0
Z
$Z_1$ $Z_2$ $Z_3$ ⋮ $Z_p$
$G(\Theta_G)$
G output
G-imputed
$\tilde{X}_1$ $\tilde{X}_2$ $\tilde{X}_3$ $\tilde{X}_4$ ⋮ $\tilde{X}_p$
+
Observed
$X_1$ ○ $X_3$ ○ ⋮ ○
=
D input
$X_1$ $\tilde{X}_2$ $X_3$ $\tilde{X}_4$ ⋮ $\tilde{X}_p$
$\delta_1$ $\delta_2$ $\delta_3$ ⋮ $\delta_q$
1 0 1 0 ⋮ 0
Covariates, fully observed
$Z_1$ $Z_2$ $Z_3$ $Z_4$ ⋮ $Z_p$
$D(\Theta_D)$
$\pi^{(v)}(X, M, Z)$
$\pi$
$\pi_1$ $\pi_2$ $\pi_3$ $\pi_4$ ⋮ $\pi_p$
D output: conditional missing probabilities

Figure 3: Diagram of the proposed algorithm.

## 3.3 Monotonicity and Convergence

We demonstrate convergence of the algorithm first by showing that it is monotone, then positing that there are ground truths for the parameters associated with $D$ and $G$. Then we show that the estimated in the algorithm convergences to these ground-truth values.

**Proposition 1.** *For random variable $X$ with missing entries, with observed section $X^{obs}$ and missing entries that are imputed by $\dot{X}(\Theta_G)$*

$$\mathcal{W}_1\big(\hat{\Theta}_G, \hat{\Theta}_D\big) := \mathcal{W}_1(\hat{\mathbf{F}}_{(obs)}, \hat{\mathbf{F}}_{(impu)}(\hat{\Theta}_G), \hat{\pi}(\hat{\Theta}_D), 1 - \hat{\pi}(\hat{\Theta}_D))$$

*is monotone decreasing with respect to gradient descent on $\mathcal{L}_G$ if:*

1. *Mild smoothness conditions: i.e. $\nabla_G \int (\cdot)dx = \int \nabla_G(\cdot)dx$*
2. *The integrals of the ECDFs of the imputed values are greater or equal than those of the observed, i.e the missing components are more skewed than observed and hence reflective of MNAR (Missing not at random)*

$$\int_0^1 |Q_{\dot{X}(\Theta_G)^{(t)}}(\tau; 1 - \hat{\pi}(\hat{X})^{(t)}|d\tau \geq \int_0^1 |Q_{X^{obs,(t)}}(\tau; \hat{\pi}(\hat{X})^{(t)})|d\tau$$

3. *There exists optimal functions $(\hat{F}_{X^{obs}}, \hat{\pi}(\hat{X}))$ and $(\hat{F}_{\dot{X}}, 1 - \hat{\pi}(\hat{X}))$ such that Assumption is satisfied elaborate on this*

*Proof.* Appendix B.1 □

Now after the likelihood $\mathcal{L}_G$ is demonstrated to be monotonically decreasing with the above assumptions, we propose that it converges to a global minimum with the ground-truth parameters $\Theta_G$ and $\Theta_D$.

**Proposition 2.** *$\mathcal{L}_G^{(t)}$ and $\mathcal{L}_D^{(t)}$ are monotone with every iteration $t$ and converge as $t$ increases:*

$$\mathcal{W}_1\big(\hat{\Theta}_D, \hat{\Theta}_G\big) \leq ... \leq \mathcal{L}_G^{(t+1)} \leq \mathcal{L}_G^{(t)}$$

*Proof.* Appendix B □

**Difference from traditional IPW or AIPW** Our approach builds on existing approaches of Seaman and White (2013) with the notable difference of applying a neural network to estimate the weights. Scharfstein et al. (1999); Li et al. (2013) describe the various problems with identifiability in regression, noting that the weighting parameter $\pi(X)$ is not possible to be identified in the response model setting. Han et al. (2019) posit a general framework for imputation using reweighting techniques. add more

## 3.4 Identifiability of Neural Network Parameters

Earlier in Section 2.1 we established general conditions for the identifiability of distribution and weighting parameters contained in $\boldsymbol{\theta}$. Now we connect these conditions specifically to neural networks. The network is parametric in that there finite, clearly defined parameters: $\boldsymbol{\Theta} = \{\Theta_D, \Theta_G\}$. $\Theta_D$ and $\Theta_G$ are the parameters of the fully connected neural networks $D$ and $G$. For each layer $l$ of either network,

$$H_G^l(X) = \begin{cases} \text{relu}(\Theta_G^{W,l} \cdot H_{l-1}(X) + \Theta_G^{b,l}) & \text{if } l > 1 \\ \text{relu}(\Theta_G^{W,1} \cdot X + \Theta_G^{b,1}) & \text{if } l = 1 \end{cases}$$

where $s(\cdot)$ is any nonlinear function. The full set of $\Theta^* = \{\Theta_w^*, \Theta_b^*\}$ comprise the parameters for $\Theta^* = \Theta_D$ or $\Theta_G$. We set both networks to have three layers (as is typical in GANs). The nodes of each layer can be chosen by the user: in the *fully connected* case, each layer has $p^2$ *weights* and $p$ *biases* (i.e. intercepts) and as such both $G$ and $D$ yield $3(p^2 + p)$ parameters. Parameter identifiability in deep neural networks has been discussed by Bona-Pellissier et al. (2023) and Phuong and Lampert (2020).Phuong and Lampert (2020) define two notions of a ReLu network and state that if these two conditions are satisfied, then the network's parameters are identifiable.

**Definition 1.** *(Phuong and Lampert (2020)) Given a set $\Omega \in \mathbb{R}^{n_K}$, if there is a $K-$layered ReLU neural network $\mathcal{N}$ with parameters $\Theta_W, \Theta_b$. If $\mathcal{N}(\Theta_W, \Theta_b)$ is*

1. *General: dense, fully connected, and each layer is full rank and whose number of neurons (e.g. parameters) per layer are non increasing,*
2. *Transparent: for any input, at least one unit (eg weight-bias combination) is active, or greater than zero*

*then the parameters $\Theta_W, \Theta_b$ uniquely characterize the network up to permutation and scaling.*

Kůrková and Kainen (1994) cite the identifiability of *shallow (one layer)* sigmoid networks, which are a one-to-one mapping in in the case of normalized data. When the *outermost layer* of the networks $G$ and $D$ are sigmoid, the ReLU identifiability result from Phuong and Lampert (2020), the combined Relu-sigmoid networks are identifiable up to scaling and permutation for both networks $G$ and $D$ if each layer is full rank. In practice, the SGD estimates for weight matrices of $\Theta_G$ and $\Theta_D$ estimated using the *GRWAIN* algorithm (Section 3.2) are always full rank at every layer.

Now we connect the conditions in Section 2.1 posit identifiability for general parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_f, \boldsymbol{\theta}_\pi)$ with neural network parameters. We specify the identifiability conditions when the distributiion parameters $\boldsymbol{\theta}_f$ and weighting parameters $\boldsymbol{\theta}_\pi$ as the parameters of respective three-layer ReLU networks

$$\boldsymbol{\theta}_f := \boldsymbol{\Theta}_G, \boldsymbol{\theta}_\pi := \boldsymbol{\Theta}_D.$$

Patra and Sen (2012) posit a general case for mixture CDFs comprising of **known** *background* CDF $F_b$ and unknown signal CDF $F_s$, constant $\alpha \in [0, 1]$. $F(x)$ can be expressed as a linear combination of $F_s(x)$ and $F_b(x)$ with linear combination $\alpha \in (0, 1)$ and $1 - \alpha$. $F(x) = \alpha F_s(x) + (1 - \alpha) F_b(x)$ with linear combination

$\alpha \in (0, 1)$ and $1 - \alpha$. A natural candidate for *mixing parameter* $\alpha$ in the imputation context is the rate of observedness (or *bias parameter* Scharfstein et al. (1999)). As such, when $\alpha$ is $\pi(X)$ , for CDFs $F^{obs}$ and $F^{mis}$ corresponding to the densities $f^{obs}$ and $f^{mis}$ as defined in (1), An analogous expression for $\pi(X)$ and $1 - \pi(X)$:

$$F_X(x) = \pi(X) \cdot F_{X^{\text{obs}}}(x) + (1 - \pi(X)) \cdot F_{\dot{X}(\Theta_G)}(x) \tag{17}$$

**Lemma 2.** *(Lemma 5 from Patra and Sen (2012)) Suppose $F_{X^{obs}}$ and $F_{X^{miss}}$ are absolutely continuous CDFs with densities $f_{X^{obs}}$ and $f_{\dot{X}(\Theta_G)}$, if the support of $F_{\dot{X}(\Theta_G)}$ is strictly contained in that of $F_{X^{obs}}$, then if either $F_{X^{obs}}$ or $F_{\dot{X}(\Theta_G)}$ is identifiable, the mixture model in Equation (17) is identifiable.*

Heuristically, we do not impute any values that are outside the obsvered values $X$. This is enforced practically by the normalization procedure of observed values such that the minimum and maximum is constrained to 0 and 1. This assumption follow the Condition in Han et al. (2019) which "guarantees the identifiability" of the quantile function.

**Assumption 1.** *$Q^0_{\dot{X}(\Theta)}(\tau)$ is the unique $\tau$-th quantile for imputed values generated by $G_{\Theta_G}(\cdot)$ (Han et al. (2019)).*

**Proposition 3.** *If Assumption 1, Lemma 2 and the conditions of Phuong and Lampert (2020) hold, then if the outputs*

$$G_{\Theta_G}(\mathbf{X}) \circ D_{\Theta_D}(\mathbf{X}) \neq G_{\Theta^*_G}(\mathbf{X}) \circ D_{\Theta^*_D}(\mathbf{X})$$

*then that implies*

$$\Theta_G \neq \Theta^*_G \textbf{\textit{ and }} \Theta_D \neq \Theta^*_D$$

*where $G \circ D$ denotes the weighted empirical quantile function difference between the imputations obtained form $G_{\Theta_G}(\cdot)$ and , using the output of $D_{\Theta_D}(\cdot)$ as the weights*

$$G_{\Theta_G} \circ D_{\Theta_D}(\mathbf{X}) = \int_0^1 |Q_{\mathbf{X}^{obs}}(\tau; D_{\Theta_D}(\hat{\mathbf{X}})) - Q_{\dot{\mathbf{X}}(\Theta_G)}(\tau, 1 - D_{\Theta_D}(\hat{\mathbf{X}}))| d\tau$$

Following Conditions 2, $2'$ in Section 2.1, Combined with Assumption 1 and Lemma 2, Proposition 3 serves as a feasible condition for identifiability using the two neural networks $G$ and $D$ in the GAN setting. As a consequence of Proposition 3, one can deduce that there exist $(f_{\dot{X}}, \hat{\pi}(X))$ such that $(f_{X^{\text{obs}}}, \pi(X))$ are uniquely best approximated by $\boldsymbol{\theta}$: For a random variable $X$ with missing and observed components $X^{\text{obs}}$ and $X^{\text{msg}}$, there exist pairs $(f_{X^{\text{obs}}}(x), \pi_0(X))$ closest to $x$ as well as pairs $(f_{\dot{X}(\Theta)}(x), 1 - \pi_0(X))$, that serve as the unique minimizers of $\mathcal{W}_1$.

# 4 Numerical Experiments

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, which may be observed or unobservable. We also consider the case where the missingness are both IM and CI. The proposed method is designated for EHR, but also can be applied to normal and uniform data.

EHR are typically heavy-tailed distributions whose patterns of missingness are self-dependent (Albers et al. (2018)), but most DL methods are not catered to MNAR data (Yoon et al. (2018); Wang et al. (2021)). There are some exceptions Ma and Zhang (2021); Dai et al. (2021), but those necessitate unique and specific settings. Most imputation algorithms like GAIN assumes MAR or MCAR. 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, INSIGHT and MIMIC-III, 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 Figure 4, right). Variables such as glucose, lipase, and creatinine are commonly modeled as lognormal or generalized extreme value (GEV) distributions (Albers et al. (2018)). We simulate experiments from these assumptions by

not sure what you mean by ”latent” and ”observed patterns”

| Simulation | Variables | $\mathbf{X}$ $(n = 300, p = 100)$ | Missingness |
|---|---|---|---|
| **IM-a** | $X_j \sim \exp(.1)$ | $\mathbf{X} = \{X_j\}_p$ | $\mathbb{P}_{\text{msg}}(X_{ij}) = \frac{\log(X_{ij}+1)}{\max_i(\log(X_{ij}+1))}$ |
| **IM-b** | $X_j \sim \text{unif}(0,1)$ | $\mathbf{X} = \{X_j\}_p$ | Same as (IM-exp) |
| **IM-c** | $X_j \sim \exp(0,.01)$ | $\mathbf{X} = \{X_j\}_p$ | Same as (IM-exp) |
| **IMCI-a** | $U_j \sim \text{unif}(0,1)$ | $\mathbf{X} = \{Y_j, Z_j\}_{j=1,\dots,p}$ | $\mathbb{P}_{\text{msg}}(Z_j)$ as IM-a |
| | $\mathbf{Z} \sim \text{MVN}(0,\ 1, \rho = .25)$ | | $\mathbb{P}_{\text{msg}}(Y_j) = \mathbb{P}_{\text{msg}}(Z_j)$ |
| | $Y_j \sim 1{+}2U_j{+}Z_j + \exp(.5)$ | | |
| **IMCI-b** | $Z_j \sim \exp(.01)$ | $\mathbf{X} = \{Y_j, Z_j\}_{j=1,\dots,p}$ | |
| | $Y_j \sim 1 + 2Z_j + \exp(.1)$ | | |
| **IMCI-d** | $Z_j \sim \exp(.1)$ | $\mathbf{X} = \{Y_j, Z_j\}_{j=1,\dots,p}$ | $\mathbb{P}_{\text{msg}}(Z_j)$ same as IM-a |
| | $Y_j \sim \text{unif}(0, 10)$ | | $\mathbb{P}_{\text{msg}}(Y_j,) \propto Z_j$ only |
| **IMCI-e** | $Z_j \sim \exp(.01)$ | $\mathbf{X} = \{Y_j, Z_j\}_{j=1,\dots,p}$ | same as above |
| | $Y_j \sim \text{unif}(0, 10)$ | | |
| **IMCI-f** | $Z_j \sim \text{unif}(0, 10)$ | $\mathbf{X} = \{Y_j, Z_j\}_{j=1,\dots,p}$ | same as above |
| | $Y_j \sim \exp(.1)$ | | |
| **CI-a** | $Z_j \sim \exp(.1)$ | $\mathbf{Y} = \{Y_j\}_{j=1,\dots,p}$ | $\mathbb{P}_{\text{msg}}(Z_j)$ same as IM-a |
| | $Y_j \sim \text{unif}(0, 10)$ | | $\mathbb{P}_{\text{msg}}(Y_j,) \propto Z_j$ only |

Table 1: Description of Simulation Schemes

In each simulation, we generate either correlated or independent random variables or mixtures of random variables. We generate independent exponential and uniform variables for the IM experiments. Higher values are more likely to be missing as to be *informative* in the experiments. In the covariate imbalance (CI) cases, pairs of variables $(X_j, Y_j)_{j=1,\dots,p}$ are generated even though they are independent of each other: $(X_j, Y_j)$ are connected through having the same missingness probabilities. Both variables are only dependent on one missingness probability vector which are generated as $\log(X_j + 1)$ for $X_j$ only. Such a case is pervasive in response-variable analyses from EHR, which are often the applicable models (Scharfstein et al. (1999); Li et al. (2013)). In the IMCI cases, variables were generated in vector pairs $(Y_j, Z_j)$ for variable index $j$, such that their concatenation forms the final dataset $\mathbf{X}$. but the variables are correlated with each other. Details of distributions and missingness among the simulations are described in Table 1.

We generated pairs of uniform and exponential random variables, as well as uniform and correlated standard normals. We vary the number of observations and variables for several choices of $n$s and $p$s. Results for these varying settings are shown in Figure 5 across different methods.

if we view generator as

## 4.1 Method Comparison for Simulations

To evaluate the proposed method, we designed experiments on synthetic datasets and compared it with MICE, MIDA, not-MIWAE, and GAIN. MICE is considered state of the art for missing data; we use MICE-CART because it outperforms other options of MICE, such as MICE-RF (Wang et al. (2021)). GAIN is theoretically sound and has partly inspired the framework of GRWAIN, but it does have limitations: Wang et al. (2021) have shown that it does not impute as well as MICE. We also compare our method to 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 100 Monte Carlo replicates for each distributional specification.

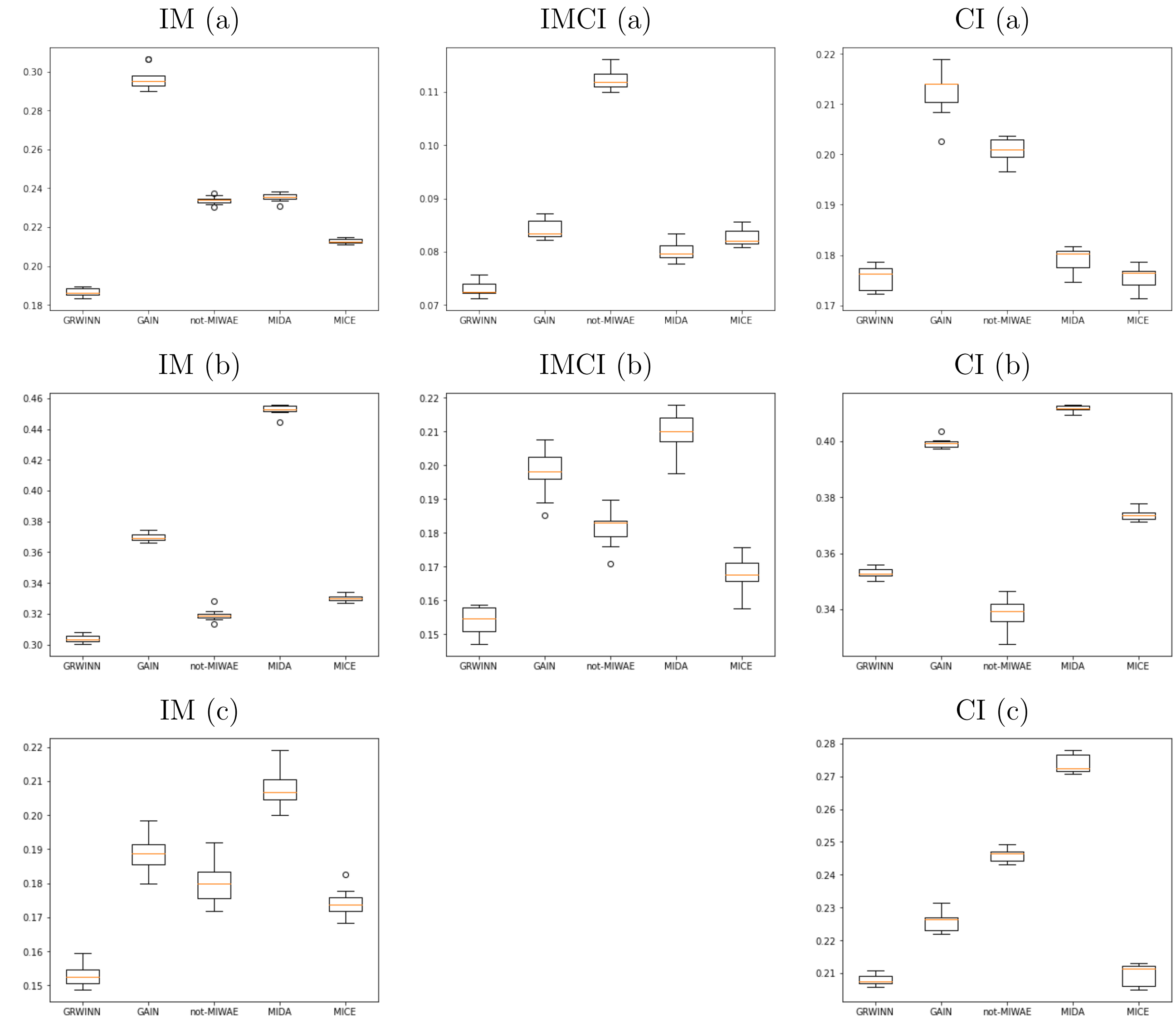
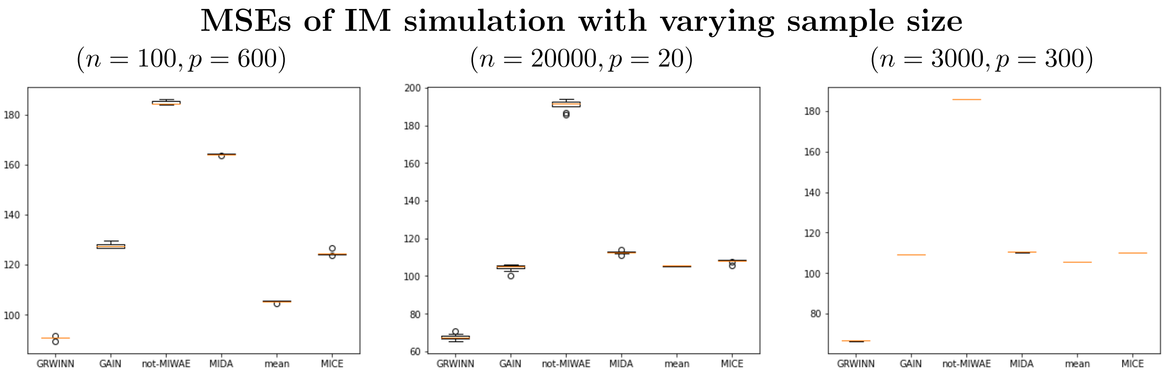


Figure 4: Simulations for *Informative Missing* (IM) hre the variables are independent exp (IM-a) and uniform (IM-b), IM with *Covariate Imbalance* (IMCI), and only Covariate Imbalance (CI), using the GRWAIN, GAIN, MIDA, MICE, as well as not-MIWAE Ipsen et al. (2020) methods. All variables except for the (standard) multivariate normals are independent. More details are found in Table 1 in the Appendix. In each scenario the number of observations $n =$ is 300 and variables $p = 100$. The simulations are normalized between 0 and 1. Boxplots show MSEs across 50 synthetic experiments for each setting.

GRWAIN outperforms other methods in most cases. For IM-a (where the missingness structure is more simplistic and closer to MAR), GRWAIN outperforms all other methods. 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. In terms of computation speed, different methods slow down in different ways as $p$ and $n$ become large. When $n$ is large and $p$ and is small (row 2 of Table in FIgure 5), GRWAIN is faster than MIDA and nMIWAE, but slower than MICE and GAIN. When $p$ is larger than $n$ (eg high dimensions low sample size), GRWAIN is the slowest, followed by MICE. When both $p$ and $n$ are large, GRWAIN is slower than MIDA, GAIN, and nMIWAE, but is faster than MICE. These methods are all feasibly computable in the operational setting and can handle datasets with millions of entries on a personal computer.



| **Computation Times** | | | | | | |
|---|---|---|---|---|---|---|
| | GRWAIN | GAIN | MIDA | nMIWAE | MICE | Mean |
| $(n = 100, p = 600)$ | 212 ± 51 | 22 ± 3 | 21± 4 | 98 ± 50 | 169 ± 50 | 0 |
| $(n = 20000, p = 20)$ | 22 ± 3 | 9 ± 1 | 48± 5 | 49 ± 6 | 5 ± 1 | 0 |
| $(n = 5000, p = 40)$ | 17± 3 | 9 ± 1 | ± 5 | ± 6 | ± 1 | 0 |
| $(n = 10000, p = 30)$ | ± 3 | ± 1 | ± 5 | ± 6 | ± 1 | 0 |
| $(n = 3000, p = 300)$ | 257 ± 28 | 30 ± 13 | 65 ± 5 | 76 ± 13 | 298 ± 70 | 0 |

Figure 5: (Top boxplots) Simulation with informative misisng (IM) with large $p$ (600) and larger $n$ (20000), and larger $n$ and $p$ (3000,300) (Bottom table) Computation times for different methods

Now we evaluate the method on simulations that may be partially missing at random (MCAR) mixed with IM. We evaluate imputation methods in two settings: one that is similar as the (*only*) IM setting described before, and one that is uniform (0,10). We apply the algorithms to the IM and MCAR datasets concurrently, but analyse them with ground-truth values separately. MCAR is induced by simply simulating 40% random (Bernoulli) missingness for all variables, while IM is dependent on the values as in (18) Under the exponential variable setting, our method does the best for both IM and MCAR variables. But under the uniform setting, MIDA, mean, and MICE perform better for the MCAR variables (Figure 6).

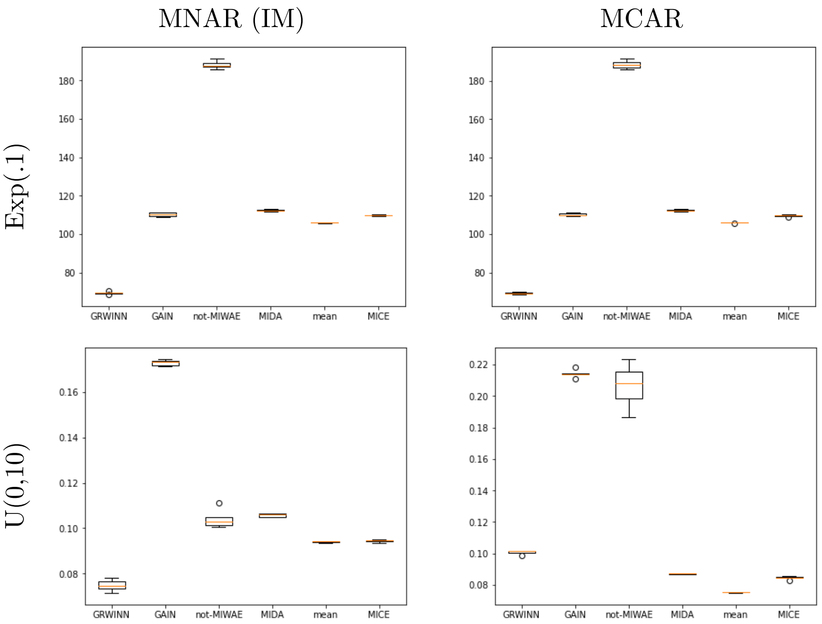


Figure 6: Simulations for *Informative Missing* (IM-exp), and for MNAR the GRWINN, GAIN, MIDA, MICE, as well as not-MIWAE Ipsen et al. (2020) methods. Simulations for *Informative Missing* (IM), MAR. The dataset is composed of both IM and MCAR samples and the method accounts for both settings, but the results are evaluated separately. Boxplots show the MSEs between the ground truth simulations and the imputed values for those that are masked-out missing from equation (18) or uniformly missing at a rate of 60%.

# 5 Case Studies and Data Analysis

We use three sources of Electronic Health Records: (1) REactions to Acute Care and Hospitalization (REACH), (2) INSIGHT clinical research network, and (3) MIMIC-III (Medical Information Mart for Intensive Care III). 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. This dataset includes symptoms, lab results, and other related information (Birk et al. (2019)). These variables provide unique information to identify modifiable environmental factors in the ED and interpersonal risk factors associated with adverse medical and psychological prognosis in acute coronary syndrome patients. The INSIGHT Clinical Research Network houses EHR data of over 12 million patients in New York City from its largest private healthcare systems (?? (INS)) . is This retrospective cohort study used de-identified EHR data from the INSIGHT Clinical Research Network (CRN), which contains EHR data of over 12 million patients in New York City.

For each of the instances of data below, we induced informative missingness (MNAR) by the following equation:

$$\mathbb{P}(X_j \text{ is missing}) = \frac{\log(\kappa_j + 1)}{\max(\log(\kappa_j + 1))} \tag{18}$$

where $\kappa_j = \frac{\text{rank}(X_j))^2}{n_j}$. For real data, we simulated 100 trials of the data masking procedure, for each variable in (18) and compared the imputation results for each imputation algorithm. Results are demonstrated in the following figures in the : boxplots of MSEs and donwstream measures are presented as there is stochasticity in the rates of missingness (even though the overall informative missingness dependency on the variables is unchanged)

Mean squared errors are typically used as the tools to assess the imputation performance (Yoon et al. (2018); Gondara and Wang (2017)). These measures are useful, but Venugopalan et al. (2019) cite the importance of evaluating imputations' impacts on the analysis that they are meant to perform. As such, for the analysis of the REACH dataset, we replicate the cox regression used in Birk et al. (2019) and use the C-index to validate the imputation's effects on model validity for *downstream analysis*.

## 5.1 Analysis of INSIGHT

The INSIGHT data collection 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 Appendix; further specificities are in Appendix F.2. The resultant data is fairly sparse; each variable is heavy tailed. This is indicative of the nature of most EHR data. INSIGHT does not have any internal benchmarks for assessing missingness; we induce $NA$s on the observed variables with the same scheme that was used in REACH. Figure 7 shows that GRWINN outperforms the other methods across a several metrics for INSIGHT: competing methods produce more bias. Across simulations and data, our proposed method reconstructs missing data more effectively both numerically and visually. Density plots show that GRWINN imputations are more similar to the original data; it does not look exactly like the *observed* nor *missing* densities but a mixture. We use MSE to judge the consensus imputation results. Point-by-point inspections of distributions show that GRWINN produces less biased and more centered imputations. In GAIN, a "hint" (subsample of mask $M$) is used in training to encourage distributional similarity between missingness and observed. In GRWINN 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 INSERT FIGUREs. Additional visual examples are found in Appendix ??.

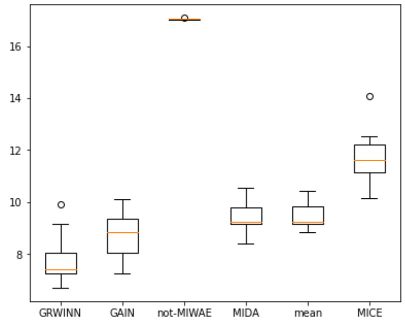


Figure 7: Mean squared errors in prediction under IM for *Insight*, with $n = 2478$ and $p = 52$ generated using missingness formula in (18).

## 5.2 Downstream Analysis of REACH

The Reactions to Acute Care and Hospitalization (ReACH) study follows 1,741 racially and socioeconomically diverse patients initially presenting to the Emergency Department with Acute Coronary symptoms.

The primary function of REACH is to analyze the association between post-traumatic stress disorder (PTSD) and cardiovascular risk. We reproduce the analysis done in Birk et al. (2019) for the time to acute coronary syndrome (ACS) which leads to major acute coronary event (MACE) or all-cause mortality (ACM). The primary aim of this study was to link time-to-MACE/ACM to PTSD, which is mostly reflected through the PCL score.

The model was a Cox Regression with the variables `"age","gender","grace", "charlson","cht","phq","pcl"` predicting `days-to-MACE` (or ACM), with the cardiac/mortality event as a binary indicator.

$$\texttt{days-to-MACE} \sim \texttt{Cox(}\ \texttt{"age","gender","grace", "charlson","cht","phq","pcl"}\ \texttt{)}$$

The model was trained on the "full" set which is comprised of 1172 observations of variables with no missing values. The concordance index (s) for this model (with the ground-truth predictors) is 0.76. Parameters $\boldsymbol{\theta}_{\text{full}}$ from the full model $\texttt{days-to-MACE} \sim \texttt{Cox}_{\boldsymbol{\theta}_{\text{full}}}(X^{\text{full}})$ are extracted and then fitted onto all of the different imputations of $X$. We use the Concordance Index (or c-index) to measure how well each model fits the data. The c-index is a metric to evaluate predictions made from survival, and is commonly used in analyses of time-to-event data.

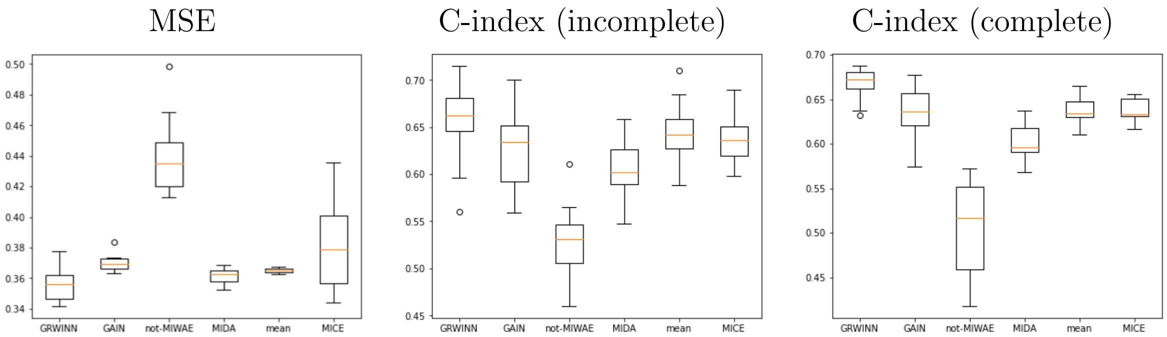
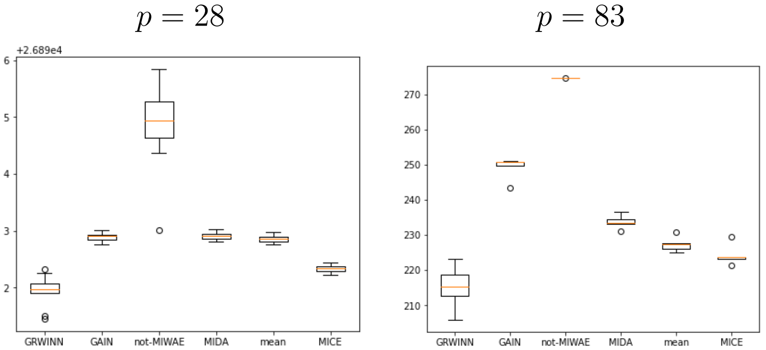


Figure 8: (Normalized) MSEs as well as Concordance scores for the *complete* ($n = 1172$) and *incomplete* ($n = 1717$) data. In both cases MNAR was generated for the cox regression analysis of predictor REACH variables, used in Birk et al. (2019)

## 5.3 Analysis of MIMIC

We also evaluated our method compared to the others on MIMIC-III (Medical Information Mart for Intensive Care III). MIMIC is a publicly available dataset that contains de-identified health data from over 50,000 patients who were admitted to critical care units in the United States. The data was collected between 2008 and 2019 and includes information such as demographics, diagnosis codes, medications, laboratory measurements, and vital signs. Numerous recent studies on EHR have used the MIMIC dataset such as Yang et al. (2019); Venugopalan et al. (2019). We focused the analysis on those who were were admitted to the *Critical Care Unit* (CCU). Within this category, the total sample size $n$ is 4066. To be consistent with that of INSIGHT, we tested our algorithm and evaluated with the others on teh *laboratory values* subset of the database. Sample laboratory variables include `Basophils, Eosinophils, Lymphocytes, Monocytes, Platelet Count, and White Blood Cells`. Prior work have noted the heavy tailed nature of some laboratory variables, and performed stabilizing transformations or removed them altogether (Yang et al. (2019)) . However, our method is specifically tailored to tackle these types of variables with potential missingness not and random.

The number of variables in the numerical *laboratory records* of MIMIC-III total $p$=92, but when the data that are nearly entirely missing are removed (i.e. when less than 10 records out of 4066 are observed) the remaining number of variables total 83. The real rate of missingness is 35%, but when the additional missingness from (18) is applied, the missingness rate nearly doubles to 63%. A further filtering step is applied for another validation set. After removing those that have more than 1500 missing, the variables $p$ is 28. We perform this sub-analysis in case the records with too many missings may skew the baselines. Imputation results for both cases are shown in Figure 9. They are calculated with the MSEs between the ground-truth *masked-out* values after missingness was induced. Computing times were also calculated. GRWINN outperforms the other methods in both settings.


$p = 28$ $p = 83$

**Computation Times**

| | GRWINN | GAIN | MIDA | nMIWAE | MICE | Mean |
|---|---|---|---|---|---|---|
| ($p = 28$) | 10.6 ± .17 | 3.2 ± .32 | 11.0 ± .04 | 12.6 ± .13 | .51 ± .09 | 0 |
| ($p = 83$) | 35.2 ± 2.1 | 8.7 ± .44 | 27.5 ± 3.1 | 27.2 ± 1.7 | 12.4 ± 11.4 | 0 |

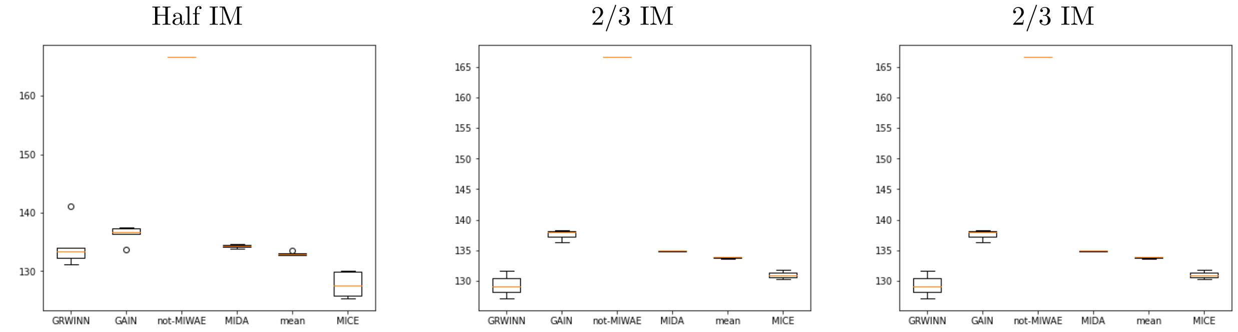
Figure 9: MIMIC mean squared errors for MIMIC-III laboratory values, **left:** normalized, **middle:** re-normalized to scale, not including nMIWAE. $n = 3235, p = 28$ (left) which comprise all the left-skewed variables that are more completely observed ($n > 1500$) .

As a comprehensive evaluation of the MIMIC dataset, we consider the root MSEs, normalized MSes (MSE divided by mean of the variable), and the 1-D Wasserstein distances for all of the sub-categories of the MIMIC CCU laboratory data. The 3 main categories for the 78 variables are ”Blood Gas”, “Hematology”, and “Chemistry”. Our method produces the lowest distances for Chemistry and Hematology values, but is slightly less effective for Blood Gas than MICE or simple mean imputation.

| | rMSE | | | | normalized MSE | | | | 1-D Wasserstein Dist. | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Blood | Chem. | Hema. | Tot. | Blood | Chem. | Hema. | Tot. | Blood | Chem. | Hema. | Tot. |
| GRWINN | 58 | 262 | 108 | 177 | 14 | 246 | 81 | 128 | 21 | 40 | 24 | 29 |
| GAIN | 59 | 288 | 130 | 197 | 16 | 295 | 103 | 156 | 23 | 54 | 38 | 41 |
| MIDA | 56 | 286 | 128 | 196 | 16 | 288 | 100 | 152 | 21 | 50 | 36 | 38 |
| MICE | 52 | 278 | 119 | 189 | 13 | 272 | 92 | 142 | 18 | 44 | 30 | 33 |
| nMIWAE | 166 | 308 | 175 | 233 | 109 | 359 | 157 | 223 | 90 | 87 | 67 | 80 |
| Mean | 52 | 284 | 118 | 191 | 14 | 282 | 92 | 146 | 20 | 49 | 34 | 36 |

Table 2: (root) MSE, normalized MSE, and Wasserstein Distance for MIMIC

Like in the simulations Section 4, we also examine two cases of MIMIC where some of the variables are IM, but others are MCAR. In the left boxplot in Figure 10, the first half of the variables are IM, the others are MCAR (missing completely at random - uniformly missing at 40%). In the right, only one third are MCAR. Our method performs better when 2/3 are MCAR, but MICE is preferable when only half are informatively missing.



(Half MCAR at 60% msg, tot 61% msg) (1/3 MCAR at 60% msg, total 62% msg) (1/3 MCAR at 80% , total 73% msg)

Figure 10: Boxplots of MSEs for imputations of MIMIC data when the missingness is half informative missing (left) and 2/3 IM (right)

# 6 Discussion

GRWINN 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 with a range of settings of missingness.

GRWAIN is designed for specific forms of EHR data, which is heavy tailed and likely to be informative missing (Albers et al. (2018)). Empirical evidence for informative missingness in EHR has been demosnrtated in Tan et al. (2023). Stanczuk et al. (2021) have discussed existing WGAN methods do not actually serve as $W_1$ minimizers , 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 in a deep learning context, but our work is novel in directly applying probability reweighing, quantile recalibration, and Wasserstein distance to a clinically pervasive problem plaguing data quality in hospitals. We have demonstrated the empirical advantages of the recalibrated Wasserstein metric $\mathcal{W}$ for purpose of imputation in informative missing EHR. We have proven some of the theoretical facets of the methods, but further work may further explore its theoretical qualities. Future directions may also explore imputation for time dependent covariates using the framework posited in this work.

# SUPPLEMENTARY MATERIAL

## A Proof of Lemma 1

We show in the following lemma that the optimal discriminator Li et al. (2017) (i.e. from Li 2018 et al.) is the probability of data $X$ *being observed*

**Lemma 3.** *For fixed generator function $G$, the optimal $D$ discriminator $D^*(\hat{X})$ is given by*

$$
\begin{aligned}
D^*(\hat{X}, \Theta_D) &= \frac{\mathbb{P}(\hat{X}, M=1)}{\mathbb{P}(\hat{X}, M=1) + \mathbb{P}(\hat{X}, M=0)} \\
&= \frac{\mathbb{P}(X^{obs}, M=1)}{\mathbb{P}(X^{obs}, M=1) + \mathbb{P}(G_{\Theta_G}^{(t-1)}(Z), M=0)} \\
&= \mathbb{E}(\pi | X, \Theta_G^{(t-1)})
\end{aligned}
$$

here $X, M, \pi$ are vectors, but can extend this expression for each $i$

*Proof.*

$$D^*(\hat{X}, \Theta_D)_i = \mathbb{E}(\pi_i | X_i, \Theta_G^{t-1})$$

$$\max_{\Theta_D} \mathcal{L}_D(X, M, \Theta_D) = \max_{\Theta_D} \mathbb{E}_{\hat{X}, M}\left[\sum_{i=1} (M_i \log(\hat{\pi}_i^{\text{obs}}) + (1 - M_i)\log(1 - \hat{\pi}_i^{\text{msg}}))\right]$$

We can rewrite the expression as follows.

$$
\begin{aligned}
&\mathbb{E}_{\hat{X}, M}\left[\sum_{i=1} (M_i \log(\hat{\pi}_i(\hat{X})) + (1 - M_i)\log(1 - \hat{\pi}_i(\hat{X})))\right] \\
&= \mathbb{E}_{X \sim f^{obs}, M}\Big[M_i \log(D(X))\Big] + \mathbb{E}_{X \sim f^{impu}, M}\Big[(1 - M_i)\log(1 - D(G(X, Z)))\Big] \\
= &\int_{\mathcal{X}} \sum_{i:M_i=1} \log(D(X|\Theta_D)_i))\mathbb{P}(X, M)dX + \int_{\mathcal{G}(X)} \sum_{i:M_i=0} \log(1 - D(G(X, Z))|\Theta_D))\mathbb{P}(G(Z), M)dG(X) \\
(*) &= \int_{\hat{\mathcal{X}}} \sum_{i=1}^{n} \log(\hat{\pi}(\hat{\Theta}_D)_i \mathbb{P}(\hat{X}_i, M_i = 1)d\hat{X} + \int_{\hat{\mathcal{X}}} \sum_{i=1}^{n} \log(1 - \hat{\pi}(\hat{\Theta}_D)_i))\mathbb{P}(\hat{X}_i, M_i = 0)dX
\end{aligned}
$$

Then differentiating with respect to $D_i$, and setting to 0, it is known for logistic functions that the maximum of $a\log(x) + b\log(1-x)$ w.r.t $x$ is $a/(a+b)$ we obtain the optimal discriminator as in (13). Let $D_i^*$ be the optimal discriminator,

$$\hat{\pi}_i(\hat{X}) = \frac{\mathbb{P}(\hat{X}_i, M_i = 1)}{\mathbb{P}(\hat{X}_i, M_i = 1) + \mathbb{P}(\hat{X}_i, M_i = 0)}$$

□

# B Monotoncity and Convergence of Algorithm

To prove convergence of the algorithe, we first demonstrate that the optimal discriminator (which has a closed form) is monotone with the generator, then we prove that the generator is monotone

**Lemma 4.** *Optimal Discriminator $D^*(\hat{X}|\Theta_D)$ is Monotone with respect to $\mathcal{L}_G$ using stochastic gradient descent*

$$\begin{aligned} D^*_\Theta(\hat{\mathbf{X}}|\Theta_G^{(t-1)}) &= \frac{P(\hat{\mathbf{X}}, \mathbf{M} = 1|\Theta_G^{(t-1)})}{P(\hat{\mathbf{X}}, \mathbf{M} = 1|\Theta_G^{(t-1)}) + P(\hat{\mathbf{X}}, \mathbf{M} = 0|\Theta_G^{(t-1)})} \\ &:= \hat{\pi}(\hat{\mathbf{X}}|\Theta_G^{(t-1)}) \end{aligned} \tag{19}$$

*Proof.* Assume $\mathbf{X} = X$ ($\mathbf{X}$ is a single vector). Plug in optimal $D^*$ into the $D$ loss.

$$\mathcal{L}_D(\Theta_D) = \mathbb{E}_{\hat{X},M}\left[\sum_{i=1}(M_i \log(\hat{\pi}_i(\hat{X})) + (1 - M_i)\log(1 - \hat{\pi}_i(\hat{X})))\right]$$

The $\mathcal{L}_D$ criterion, conditional on $M$ (as in the $G$ criterion) is: at optimal discriminator $D^*(\Theta_D|\Theta_)$ *given* the generator at iteration $t$:

$$\hat{\pi}(\hat{X})_i = D^*(X_i) = \frac{\mathbb{P}(X_i, M_i = 1)}{\mathbb{P}(G(X^{obs}, Z)_i, M_i = 0) + \mathbb{P}(X_i, M_i = 1)}$$

$$\begin{aligned} \mathcal{L}_D &= \mathbb{E}_{\hat{X},M}\left[\sum_{i=1}(M_i \log(\hat{\pi}_i(\hat{X})) + (1 - M_i)\log(1 - \hat{\pi}_i(\hat{X})))\right] \\ &= \sum_{i \text{ is observed}} \log\left(\frac{\mathbb{P}(X_i, M_i = 1)}{\mathbb{P}(G(X^{obs}, Z)_i, M_i = 0) + \mathbb{P}(X_i, M_i = 1)}\right) \\ &+ \sum_{i \text{ is missing}} \log\left(1 - \frac{\mathbb{P}(X_i, M_i = 1)}{\mathbb{P}(G(X^{obs}, Z)_i, M_i = 0) + \mathbb{P}(X_i, M_i = 1)}\right) \end{aligned}$$

then taking derivative w.r.t $G$ at $i$ missing:

$$\begin{aligned} \nabla_G \log\left(1 - \hat{\pi}_i^{\text{msg}}(\hat{X})\right) &= (1 - M_i)\nabla_G \log\left(\frac{\mathbb{P}(G(X^{obs}, Z)_i, M_i = 0)}{\mathbb{P}(G(X^{obs}, Z)_i, M_i = 0) + \mathbb{P}(X_i, M_i = 1)}\right) \\ &= (1 - M_i)\frac{\mathbb{P}(X_i, M_i = 1) \cdot \mathbb{P}'(G(X^{obs}, Z)_i, M_i = 0)}{\mathbb{P}(X_i, M_i = 1) \cdot \mathbb{P}(G(X^{obs}, Z)_i, M_i = 0) + \mathbb{P}(G(X^{obs}, Z)_i, M_i = 0)^2} \\ &\propto \mathbb{P}'(G(X^{obs}, Z)_i, M_i = 0) \end{aligned}$$

Therefore, for each component $i$ of $\mathcal{L}_D(\hat{X}, M, \Theta_D)$, that is missing:

$$\begin{aligned} \nabla_G \mathcal{L}_D(\hat{X}, M, \Theta_D)\Big|_{i \text{ is missing}} &= \nabla_G \mathbb{E}_{\hat{X},M}\left[(M_i \log(\hat{\pi}_i(\hat{X})) + (1 - M_i)\log(1 - \hat{\pi}_i(\hat{X})))\right] \\ &= \nabla_G \mathbb{E}_{G(X^{\text{obs}}, Z), M}\left[(M_i \log(\hat{\pi}_i(X^{\text{obs}})) + (1 - M_i)\log(1 - \hat{\pi}_i(G(X^{\text{obs}}, Z))))\right] \\ &\propto (1 - M_i) \propto \mathbb{P}'(G(X^{obs}, Z)_i, M_i = 0) \end{aligned}$$

So $L_D$ at a missing entry $i$ is *monotonic with* $G$: if $\nabla_G \mathbb{P}'(G(X^{obs}, Z)_i, M_i = 0)$ then $\nabla_G \{L_D\}_i > 0$. $\square$

$$\min_{\Theta_G} \mathcal{L}_G(X, M, \Theta_G) = \min_{\Theta_G} \mathcal{W}(X, G_{\Theta_G}(X, Z)).$$

In the following sections for simplicity of notation we write $\hat{F}_{G(X,Z)}$ simply as $\hat{F}_G$. Criterion $\mathcal{W}^{(k)}$ is monotone if

$$\mathcal{W}^{(k)} \leq \mathcal{W}^{(k-1)}$$

since each

$$\mathcal{W}^{(k)} = \mathcal{W}^{(k-1)} + \nabla_G(\mathcal{W}^{(k-1)})\delta$$

for some $\delta > 0$, it suffices to prove $\nabla_G(\mathcal{W}^{(k-1)}) < 0$.

## B.1 $\nabla G$ is monotone

**Proposition 4.** *For random variable $X$ with missing entries, with observed section $X^{obs}$ and missing entries that are imputed by $\dot{X}(\Theta_G)$*

$$\mathcal{W}_1\big(\hat{\Theta}_G, \hat{\Theta}_D\big) := \mathcal{W}_1(\hat{\mathbf{F}}_{(obs)}, \hat{\mathbf{F}}_{(impu)}(\hat{\Theta}_G), \hat{\pi}(\hat{\Theta}_D), 1 - \hat{\pi}(\hat{\Theta}_D))$$

*is monotone decreasing with respect to gradient descent on $\mathcal{L}_G$ if:*

1. *Mild smoothness conditions: i.e. $\nabla_G \int (\cdot) dx = \int \nabla_G(\cdot) dx$*
2. *The integrals of the ECDFs of the imputed values are always greater than $t$, i.e the missing components are more skewed than observed and hence reflective of MNAR*
$$\int_0^1 |Q_{\dot{X}(\Theta_G)^{(t)}}(\tau; 1 - \hat{\pi}(\hat{X})^{(t)}| d\tau > \int_0^1 |Q_{X^{obs,(t)}}(\tau; \hat{\pi}(\hat{X})^{(t)})| d\tau$$
3. *There exists optimal functions $(\hat{F}_{X^{obs}}, \hat{\pi}(\hat{X}))$ and $(\hat{F}_{\dot{X}}, 1 - \hat{\pi}(\hat{X}))$ such that Assumption is satisfied*

*Proof.* First we note the *inversion theorem* for quantiles to CDFs

$$\int_0^1 |\widetilde{Q}_{X^{obs}}(\tau; \hat{\pi}(\hat{X})^{(t)}) - \widetilde{Q}_{\dot{X}^{(t)}(\Theta_G)}(\tau; 1 - \hat{\pi}(\hat{X})^{(t)}| d\tau = \int \Big| \tilde{F}_{\dot{X}(\Theta_G)}(x; 1 - \hat{\pi}(X)) - \tilde{F}_{X^{obs}}(x; \hat{\pi}(X)) \Big| dx$$

Using the ECDF form, we can first move the derivative inside:

$$\nabla_G \int \Big| \tilde{F}_{\dot{X}(\Theta_G)}(x; 1 - \hat{\pi}(X)^{(t)}) - \tilde{F}_{X^{obs}}(x; \hat{\pi}(\hat{X})^{(t)}) \Big| dx = \int \nabla_G \Big| \tilde{F}_{\dot{X}(\Theta_G)}(x; 1 - \hat{\pi}(X)^{(t)}) - \tilde{F}_{X^{obs}}(x; \hat{\pi}(X)^{(t)}) \Big| dx \tag{20}$$

Note for *fixed* $\hat{\pi}(X)$ at a given iteration $t$,

$$\nabla_G^{(t)} \mathcal{L}_{G^{(t)}} = \nabla_{G^{(t)}} \mathcal{W}(\Theta_G^t) \leq 0$$

will be negative because gradient descent minimization will select for the parameteras to minimize the objective function. What is not gruaranteed, however, is whether $\mathcal{L}_G^{(t)}(\hat{\boldsymbol{\pi}})$ will be monotone with the imputed portion of the integral. For simplicity of notation, and referring back to the "generalized form" in (7)

$$\hat{\omega}_i^{miss} = \frac{1/(1-\hat{\pi}(\hat{X})_i)}{\sum_{j:j \text{ is missing}}^{n_0} 1/1-\hat{\pi}(\hat{X})_j}\Bigg|_{i \text{ is missing}} \tag{21}$$

We recall the definition $\tilde{F}_X(x;\boldsymbol{\omega})$ for any normalized weight probability vector $\boldsymbol{\omega}$ such that its components add to 1. Then using the *normalized inverse of* $\pi(X)$

$$\tilde{F}_{X^{obs}}(x;\hat{\pi}(\hat{X})) = \sum_{i=1}^{n_1} \frac{1/\hat{\pi}(\hat{X})_i}{\sum_{j=1}^{n_1} 1/\hat{\pi}(\hat{X})_j} I\{X_i \le x\}.$$

$$\tilde{F}_{\dot{X}(\Theta_G)}(x;1-\hat{\pi}(\hat{X})) = \sum_{i=1}^{n} \frac{1/(1-\hat{\pi}(\hat{X})_i)}{\sum_{j=1}^{n_0} 1/1-\hat{\pi}(\hat{X})_j} I\{\dot{X}(\Theta_G)_i \le x\}.$$

The gradient of the weighted imputation ECDF is:

$$\begin{aligned}\nabla_G^{(t)} \tilde{F}_{\dot{X}(\Theta_G)}(x;1-\hat{\pi}(X)^{(t)}) &= \nabla_{G^{(t)}} \sum_{i:i \text{ is missing}}^{n_0} \hat{\omega}_i^{miss} I(G_{\Theta_G}^{(t)}(X^{obs},Z)_i \le x) \\ &= \sum_{i:i \text{ is missing}}^{n_0} \hat{\omega}_i^{miss} \Big( -\delta(x - G_{\Theta_G}^{(t)}(X^{obs},Z)_i) \cdot |J_{\Theta_G}| \Big)\end{aligned}$$

where $\delta()$ is the dirac function. For fixed weights $1-\hat{\pi}_i$ (obtained in $D^{(t-1)}$), then for Jacobian $J_{\Theta_G}$,

$$\nabla_G^{(t)} \int_{\mathcal{X}} \tilde{F}_{\dot{X}(\Theta_G)}(x;1-\hat{\pi}(X)^{(t)})dx = \int_{\mathcal{X}} \sum_{i:i \text{ is missing}}^{n_0} \hat{\omega}_i^{miss} \Big( -\delta(x - G_{\Theta_G}^{(t)}(X^{obs},Z)_i) \cdot |J_{\Theta_G}| dx \Big) \tag{22}$$

$$\le 0 \tag{23}$$

by the properties of the dirac integral. Then, since the estimate $\hat{\pi}(X)$ at time $t+1$

$$\hat{\pi}^{t+1}(\hat{X}) = D^{*,t}(\Theta_G^t) = \frac{P(X, M=1)}{P(X,M=1) + P(G^t, M=0)}$$

$$\hat{\pi}_i(\hat{X}) = D^*(X_i) = \frac{\mathbb{P}(X_i, M_i = 1)}{\mathbb{P}(G(X^{obs},Z)_i, M_i = 0) + \mathbb{P}(X_i, M_i = 1)}$$

From Lemma 4 From Lemma 4, $\hat{\pi} = D^*$ is monotone with $G$

$$\begin{aligned}\int_{\mathcal{X}} \tilde{F}_{\dot{X}(\Theta_G)}^{(t+1)}(x;1-\hat{\pi}(X)^{(t+1)})dx &\le (a) \int_{\mathcal{X}} \tilde{F}_{\dot{X}(\Theta_G)}^{(t)}(x;1-\hat{\pi}(X)^{(t+1)}) \\ &\le (b) \int_{\mathcal{X}} \tilde{F}_{\dot{X}(\Theta_G)}^{(t)}(x;1-\hat{\pi}(X)^{(t)})\end{aligned}$$

(a): From (22) $\nabla_{G^t} \int_{\mathbb{R}} \tilde{F}_{\dot{X}(\Theta_G)}(x;1-\hat{\pi}(X)^{(t)})dx \le 0$ then $\int \hat{F}_G^{t+1} \le \int \hat{F}_G^t$ from gradient descent, so for fixed weights the relation holds

(b): Lemma 4 states $D$ is monotone with $G$

So after we move the gradient operator inside in (20)

$$\begin{aligned}
&\int \nabla_G \left| \tilde F_{\dot X(\Theta_G)}(x;1-\hat\pi(X)^{(t)}) - \tilde F_{X^{obs}}(x;\hat\pi(X)^{(t)}) \right| dx \\
&\quad = \int \frac{\tilde F_{\dot X(\Theta_G)}(x;1-\hat\pi(X)^{(t)}) - \tilde F_{X^{obs}}(x;\hat\pi(X)^{(t)})}{\left|\tilde F_{\dot X(\Theta_G)}(x;1-\hat\pi(X)^{(t)}) - \tilde F_{X^{obs}}(x;\hat\pi(X)^{(t)})\right|} \cdot \nabla_G \left( \tilde F_{\dot X(\Theta_G)}(x;1-\hat\pi(X)^{(t)}) - \tilde F_{X^{obs}}(x;\hat\pi(X)^{(t)}) \right) dx \\
&\quad \propto \int \nabla_G \tilde F_{\dot X(\Theta_G)}(x;1-\hat\pi(X)^{(t)}) dx
\end{aligned}$$

by condition 2 of Proposition 4, gradient descent of $\mathcal{W}$ is monotone. A similar argument in Ipsen et al. (2020) will show that it will converge to the true objective function.

□

- For The sum of the integrals of the normalized data , the missing part has to be greater than observed
- In orde rto be more speicfic, We have to know, for CI case, which X's are correlated with y

# C Consistency of $G$

**Lemma 5.**

$$\tilde F_{\dot X(\widehat\Theta_G)}(x;1-\hat\pi(\hat X)) \to^p F_{\dot X(\Theta_G^0)}(x;1-\pi(\hat X))$$

*where $\dot X$ is the ground truth $G^0(\cdot)$ such that $G^0(X) = \dot X$*

*Proof.* By fixing the estimate of $\hat\pi(\hat X)$ as a constant (from the previous step of $D$-optimization), and using the property of the (weighted) ECDF that

$$\mathbb{E}_{\dot X(\Theta_G)}\left[\tilde F_{\dot X(\Theta_G)}(x;1-\hat\pi(\hat X))\right] = F_{\dot X(\Theta_G)}(x;1-\hat\pi(X))$$

where $\hat\pi_i :=$ are constants, using convergence from above (G convergence), so suffices to show that

$$\begin{aligned}
\tilde F_{\dot X(\Theta_G)}(x;1-\hat\pi(\hat X)) &= \sum_{i:i \text{ is missing}}^{n} \frac{1/(1-\hat\pi(\hat X)_i)}{\sum_{j=1}^{n_0} 1/1-\hat\pi(\hat X)_j} I\{\dot X(\Theta_G)_i \leq x\} \\
&= \sum_{i=1}^{n} \hat T_i
\end{aligned}$$

here each $\hat T_i$ is contained in upper and lower bounds as $1-\hat\pi(X)_i$ is bounded away from 0 and 1, so we define $\hat T_i$ as

$$T_i := \widehat\omega_i^{(miss)}(1-M_i) I\{\dot X(\Theta_G)_i \leq x\}$$

where $\omega_i^{(miss)}$ is defined as in the previous equation (21), and where $\hat\pi(X)_i = D^*(Y)$, treated as fixed with respect to $\Theta_G$ , where $Y = G_{\Theta_G}^{(t-1)}(Z)$ from the previous iteration.

$$\mathbb{E}T_i = \omega_i^{(miss)}(1-M_i) F_{\dot X(\Theta_G)}(x;1-\pi(X))$$

Hoeffding inequality applies if , for $T = \sum_i T_i$

1. Each $T_i \in [a_i, b_i]$, because $\hat{\pi}_i \in (\epsilon, 1-\epsilon)$ by Condition 2.1 $-\infty < T_i < \infty$ and as so $-\infty < \min(T_i)$ and $\max(T_i) < \infty$

$$T_i := \widehat{\omega}_i^{(miss)}(1 - M_i)I\{\dot{X}(\Theta_G)_i \leq x\}$$

2. $T_i, T_j$ independent as each $\hat{\pi}_i$ fixed constant, and by definition of ECDFs each indicator is independent

$$\mathbb{P}(|\sum_{i=1}^{n} T_i - \mathbb{E}\sum_i T_i| > \epsilon) \leq \exp(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(\max(T_i) - \min(T_i))^2})$$

as such the probability $\downarrow 0$ as $n \uparrow \infty$ so by Hoeffding inequality.

$$\mathbb{P}\left(\left|\tilde{F}_{\dot{X}(\Theta_G)}(x; 1 - \hat{\pi}(\hat{X})) - F_{\dot{X}(\Theta_G)}(x; 1 - \hat{\pi}(\hat{X}))\right| \geq \epsilon\right) \to 0$$

## C.1 Consistency of $D$

Want to show

$$\hat{D}_{\Theta_D}(G_{\Theta_G}(X, Z)) \to \pi(X)$$

Let $\hat{\pi}(X)$ be the optimal $D^*$ given $G$ i.e.

$$\begin{aligned}\hat{\pi}(X_i) &= \mathbb{E}[\pi(X_i)|\Theta_G] \\ &= \frac{\mathbb{P}(X_i, M_i = 1)}{\mathbb{P}(X_i, M_i = 1) + \mathbb{P}(G_{\Theta_G}(X), M = 0)}\end{aligned}$$

for single vector $X$, with mask vector $M$ (indices $i$). the estimate of this is from the output of $D$. Sample estimate $\hat{M}_i$ is estimate of Bernoulli.

$$\hat{\pi}(\hat{X})_i = D^*(X_i) = \frac{\mathbb{P}(X_i, M_i = 1)}{\mathbb{P}(G(X^{obs}, Z)_i, M_i = 0) + \mathbb{P}(X_i, M_i = 1)}$$

$$M_i \sim \text{Bern}(\pi(X_i))$$

So, since from previous assertions that (1) $D^*$ is a function of $G$ from previous iteration $t-1$. By continuous mapping theorem, $D(G^{t-1})_i = \hat{\pi}_i$ also converges in probability $\mathbb{P}(|\hat{\pi}_i - \pi_i| > \epsilon) \to 0$ □

# D Wasserstein GAN

**Wasserstein Distance:** typically $X$ represents the observed data and $Y$ the generated (latent) estimations. Another definition of Wasserstein distance $W_1$ between two distributions $X$ and $Y$ is expressed as the integral of the difference of CDFs

$$W_1(F_X, F_Y) = \int_{-\infty}^{\infty} \left|F_X(x) - F_Y(x)\right| dx \tag{24}$$

The corresponding sample Wasserstein distance is similarly defined,

Here we briefly describe GAN and its Wasserstein variant. Consider a $p$-dimensional space $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_p$, let $(x_{\text{data}}, x_G)$ be elements in $\mathcal{X}$ whose marginal distributions are $P_{\text{data}}, P_G$. $\gamma(x_{\text{data}}, x_G)$ indicates how much mass must be transported from $x_{\text{data}}$ to $x_G$ in order to transform the distribution $P_{\text{data}}$ *into* $P_G$. The Wasserstein-1 distance $W_1$ is the minimum cost of the *optimal transport* plan between the distributions (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:

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

# E Additional Simulation Tables

## E.1 MIMIC details

### E.1.1 Names of Laboratory Test Values

`Calculated Total CO2,Glucose,Hematocrit, Calculated,Hemoglobin,Lactate,Oxygen ,pCO2,pO2,Potassium, Whole Blood,Alanine Aminotransferase (ALT),Alkaline Phosphatase,Amylase,Anion Gap,Asparate Aminotransferase (AST),Bicarbonate,Bilirubin, Total,Calcium, Total,CK-MB Index, Creatine Kinase (CK),Creatine Kinase, MB Isoenzyme, Creatinine, Glucose,Lactate Dehydrogenase (LD),Lipase, Magnesium,Phosphate,Potassium, Thyroid Stimulating Hormone, Triglycerides,Troponin T,Urea Nitrogen, Creatinine, Urine, Sodium, Urine, Bands, Basophils, Eosinophils, Hematocrit, INR(PT), Lymphocytes, Monocytes, Platelet Count, PT, PTT, RDW, Red Blood Cells, White Blood Cells, Epithelial Cells,pH,Protein,RBC,Specific Gravity,WBC`

# F 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 11 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

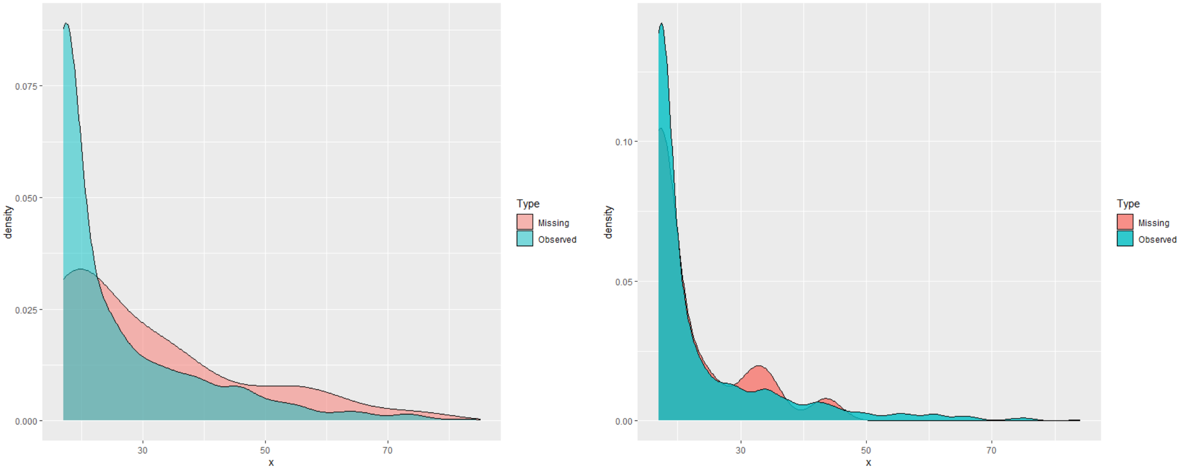


Figure 11: (left) Density plots for observed (i.e. same value in raw and cleaned datasets) and missing (i.e. retroactively filled-in in the "cleaned" dataset) PCL baseline and (right) PCL 12 months

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 **??**.

**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).

## F.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 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

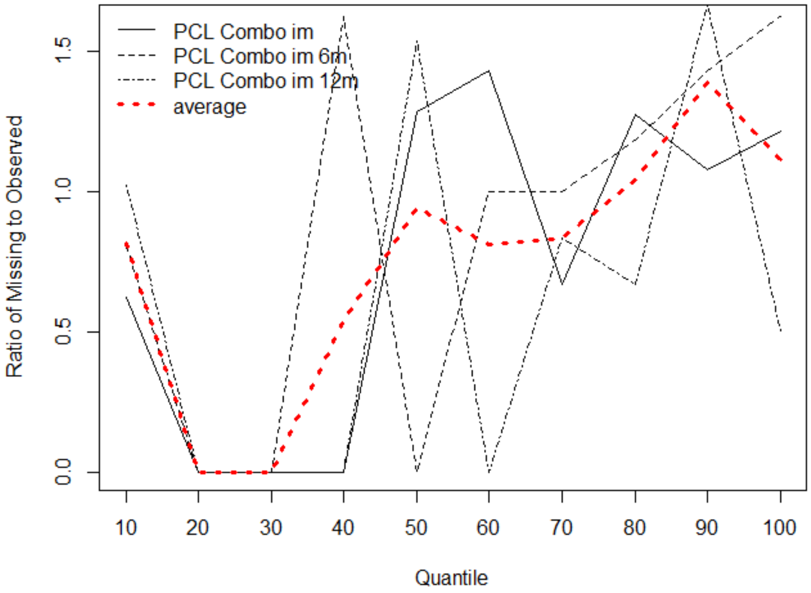


Figure 12: Ratios of the proportions of missing (i.e. $Y_j^I$) variables for PCL Combo scores at baseline, 6mo, 12mo, to those that are only observed (i.e. $Y_j^O$) , that fall within the decile levels as derived from the common dataset $Y_j^C$. The red line indicates an average of the PCL score ratios 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).

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 11).

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

where $\kappa_j = \frac{(\text{rank}(X_j))^2}{n_j}$. Missingness increase as the values get larger, with a probability of 1 as the5 values approach maximum.

## F.2 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).

# G Details for Simulation and Data Analysis

## G.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.

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

## G.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 $Y_j \sim 1 + 2X_j + \exp(.1)$.

# H Software

**Title:** Brief description.

**R-package xxxx:** R-package xxxx

**Data set xxx:** Data set xxx

# References

Insight clinical research network.

Albers, D., N. Elhadad, J. Claassen, R. Perotte, A. Goldstein, and G. Hripcsak (2018). Estimating summary statistics for electronic health record laboratory data for use in high-throughput phenotyping algorithms. *Journal of Biomedical Informatics 78*, 87–101.

Arjovsky, M., S. Chintala, and L. Bottou (2017). Wasserstein GAN.

Birk, J., I. Kronish, B. Chang, T. Cornelius, M. Abdalla, J. Schwartz, J. Duer-Hefele, A. Sullivan, and D. Edmondson* (2019, January). The Impact of Cardiac-induced Post-traumatic 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.

Bona-Pellissier, J., F. Bachoc, and F. Malgouyres (2023). Parameter identifiability of a deep feedforward relu neural network.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1983). Classification and regression trees.

Cheng, H. and Y. Wei (2018, dec). A fast imputation algorithm in quantile regression. *Comput. Stat. 33*(4), 1589–1603.

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

Friedjungová, M., D. Vašata, M. Balatsko, and M. Jiřina (2020). Missing features reconstruction using a wasserstein generative adversarial imputation network. In V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos, and J. Teixeira (Eds.), *Computational Science – ICCS 2020*, Cham, pp. 225–239. Springer International Publishing.

Gondara, L. and K. Wang (2017). Mida: Multiple imputation using denoising autoencoders.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). Generative adversarial networks.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology 60*(1), 549–576. PMID: 18652544.

Han, P., L. Kong, J. Zhao, and X. Zhou (2019). A general framework for quantile estimation with incomplete data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 81*(2), 305–333.

Ipsen, N. B., P.-A. Mattei, and J. Frellsen (2020). not-miwae: Deep generative modelling with missing not at random data.

Kůrková, V. and P. C. Kainen (1994). Functionally equivalent feedforward neural networks. *Neural Computation 6*(3), 543–558.

Li, J., A. Madry, J. Peebles, and L. Schmidt (2017). On the limitations of first-order approximation in GAN dynamics.

Li, L., C. Shen, X. Li, and J. M. Robins (2013). On weighting approaches for missing data. *Statistical Methods in Medical Research 22*(1), 14–30. PMID: 21705435.

Li, S. C.-X., B. Jiang, and B. Marlin (2019). Misgan: Learning from incomplete data with generative adversarial networks.

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

Luo, Y., X. Cai, Y. ZHANG, J. Xu, and Y. xiaojie (2018). Multivariate time series imputation with generative adversarial networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 31. Curran Associates, Inc.

Ma, C. and C. Zhang (2021). Identifiable generative models for missing not at random data imputation. *CoRR abs/2110.14708*.

Mattei, P.-A. and J. Frellsen (2019). Miwae: Deep generative modelling and imputation of incomplete data.

Ostrovski, G., W. Dabney, and R. Munos (2018). Autoregressive quantile networks for generative modeling.

Patra, R. K. and B. Sen (2012). Estimation of a two-component mixture model with applications to multiple testing.

Phuong, M. and C. H. Lampert (2020). Functional vs. parametric equivalence of relu networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association 94*(448), 1096–1120.

Seaman, S. R. and I. R. White (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research 22*(3), 278–295. PMID: 21220355.

Stanczuk, J., C. Etmann, L. M. Kreusser, and C.-B. Schönlieb (2021). Wasserstein gans work because they fail (to approximate the wasserstein distance).

Talas, L., J. G. Fennell, K. Kjernsmo, I. C. Cuthill, N. E. Scott-Samuel, and R. J. Baddeley (2020). Camogan: Evolving optimum camouflage with generative adversarial networks. *Methods in Ecology and Evolution 11*(2), 240–247.

Tan, A. L., E. J. Getzen, M. R. Hutch, Z. H. Strasser, A. Gutiérrez-Sacristán, T. T. Le, A. Dagliati, M. Morris, D. A. Hanauer, B. Moal, C.-L. Bonzel, W. Yuan, L. Chiudinelli, P. Das, H. G. Zhang, B. J. Aronow, P. Avillach, G. Brat, T. Cai, C. Hong, W. G. La Cava, H. Hooi Will Loh, Y. Luo, S. N. Murphy, K. Yuan Hgiam, G. S. Omenn, L. P. Patel, M. Jebathilagam Samayamuthu, E. R. Shriver, Z. Shakeri Hossein Abad, B. W. Tan, S. Visweswaran, X. Wang, G. M. Weber, Z. Xia, B. Verdy, Q. Long, D. L. Mowery, and J. H. Holmes (2023). Informative missingness: What can we learn from patterns in missing laboratory data in the electronic health record? *Journal of Biomedical Informatics 139*, 104306.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software 45*(3), 1–67.

Venugopalan, J., N. Chanani, K. Maher, and M. D. Wang (2019). Novel data imputation for multiple types of missing data in intensive care units. *IEEE Journal of Biomedical and Health Informatics 23*(3), 1243–1250.

Villani, C. (2008). Optimal transport: Old and new.

Wang, Z., O. Akande, J. Poulos, and F. Li (2021). Are deep learning models superior for missing data imputation in large surveys? evidence from an empirical comparison.

Wei, Y., Y. Ma, and R. J. Carroll (2012). Multiple imputation in quantile regression. *Biometrika 99*(2), 423–438.

Xie, Y. and B. Zhang (2017, 01). Empirical likelihood in nonignorable covariate-missing data problems. *The International Journal of Biostatistics 13*.

Yang, X., Y. J. Kim, F. Khoshnevisan, Y. Zhang, and M. Chi (2019). Missing data imputation for mimic-iii using matrix decomposition. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 1–3.

Yang, Y., Z. Wu, V. Tresp, and P. A. Fasching (2019, jun). Categorical EHR imputation with generative adversarial nets. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE.

Yoon, J., J. Jordon, and M. van der Schaar (2018). Gain: Missing data imputation using generative adversarial nets.

Yoon, J., W. R. Zame, A. Banerjee, M. Cadeiras, A. M. Alaa, and M. van der Schaar (2018, 03). Personalized survival predictions via trees of predictors: An application to cardiac transplantation. *PLOS ONE 13*(3), 1–19.

Yuan, X. and X. Dong (2019). Weighted empirical likelihood for quantile regression with non ignorable missing covariates. *Communications in Statistics - Theory and Methods 48*(12), 3068–3084.