INTEGRATED PU LEARNING AND DEEP GENERATIVE VAES FOR INDICATION-DRIVEN MISSINGNESS TO IMPROVE MNAR IMPUTATION AND PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in deep generative modeling have improved imputation for missing not at random (MNAR) data. However, jointly improving both imputation quality and downstream prediction remains challenging. In this paper, we propose to use the positive-unlabeled (PU) learning to deal with indication-driven missingness to leverage partially observed measurement-intent variable to improve MNAR imputation and downstream prediction. We first use a PU adversarial network to learn probabilities of unknown positives of intention to measure among all the missing cases, which then guide the training of variational autoencoders (VAE) that jointly model observed data and missingness indicators. Experiments on synthetic and real-world data show that our proposed framework has superior performance for both imputation and downstream prediction tasks. These results show that explicitly modeling missingness indicators integrated with additional PU measurement intent information can improve MNAR imputation and downstream predictions under various scenarios and assumptions.

1 INTRODUCTION

Missing data are common in tabular datasets and can arise for a variety of reasons, including data collection errors, non-response, or a variable was not intended to be collected (Little & Rhemtulla, 2013; Emmanuel et al., 2021). Appropriate handling of missing data is therefore important, as improper handling can reduce the reliability of statistical analyses and impair the predictive performance of models (Nijman et al., 2022; Schafer & Graham, 2002).

A broad range of methods has been developed for missing-data imputation. Early approaches include simple imputation based on summary statistics, as well as statistical and machine learning methods such as Multiple Imputation by Chained Equations (MICE) (Buuren & Groothuis-Oudshoorn, 2011), k-NN (Murti et al., 2019), Bayesian models (Ma & Chen, 2018), and missForest (Stekhoven & Bühlmann, 2012). Deep generative models have also been used for imputation, including Variational AutoEncoders (VAEs) (Mattei & Frellsen, 2019; Ivanov et al., 2019) and Generative Adversarial Networks (GAN) (Yoon et al., 2018; Li et al., 2019; Dai et al., 2021). When data are missing not at random (MNAR), the probability that a value is missing depends on the unobserved data and therefore needs to be explicitly modeled (Rubin, 1976; Little & Rubin, 2019). Representative deep generative approaches, such as not-MIWAE (Ipsen et al., 2021) and GINA (Ma & Zhang, 2021), jointly model the data distribution and missingness mechanism. In these methods, missingness is typically represented by a binary mask indicating whether a value is observed.

However, a binary missingness mask tells us whether a value is available, but not whether that value was intended to be measured. As a result, variables that were not intended to be measured and those that were intended to be measured but ultimately remained missing are encoded identically. This distinction is particularly relevant in EHRs, where laboratory testing is often driven by clinical need and measurement itself can be informative of patient health status (Agniel et al., 2018; Tan et al., 2023). Sisk et al. (2021) described related phenomena as *informative presence*, where whether a variable is measured conveys information about the subject's condition, and *informative observation*, where the timing or frequency of repeated measurements conveys information. Sun et al. (2024) used

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the term *informatively collected laboratory data*. However, these studies did not provide a standard term or formal definition for this type of item-level missingness.

We introduce the term *indication-driven missingness* to describe item-level missingness arising when measurement is selectively initiated by subject-specific indications. Unlike planned missingness, which is imposed prospectively by study design, indication-driven missingness arises from a decision process in which the presence or absence of symptoms, signs, risk factors, preliminary findings, or professional judgment determines whether a variable is measured. To our knowledge, this is the first formulation of *indication-driven missingness* as a new type of missing data through a partially observed variable and PU learning. *Indication-driven missingness* may arise under either MAR or MNAR; in our real data example, it is plausibly MNAR because the decision to measure may depend on subject-specific information that is not fully observed and may therefore be informative about both the missing value and downstream outcomes.

We further formulate *indication-driven missingness* as a Positive-Unlabeled (PU) learning problem (Kiryo et al., 2017; Elkan & Noto, 2008; Kato et al., 2019), because it is only partially observed in practice. Observed evidence of an intention to measure defines the *known positive* set, whereas the remaining instances are *unlabeled* and may contain either unrecorded positive intent or true negatives. We can then use and model a new *measurement-intent* indicator when the *indication-driven missingness* is present. This formulation is analogous to other fields, such as implicit feedback in recommender systems (Wang et al., 2026), click feedback in learning-to-rank (Radlinski & Joachims, 2005), and selective labels in human decision-making (Lakkaraju et al., 2017).

This ambiguity in *measurement-intent* creates an additional challenge for MNAR imputation. Measurement intent may contain information beyond the binary missingness mask, including information relevant to downstream outcomes (Agniel et al., 2018; Che et al., 2018; Lipton et al., 2016). Because imputing a missing value does not necessarily recover whether that variable was intended to be measured, collapsing unlabeled positives and true negatives within the unlabeled set may discard information useful for both imputation and prediction. PU learning provides a natural way to recover this latent measurement-intent information and can integrate *indication-driven missingness* into both tasks.

This paper introduces PU-VAE-MNAR, a deep generative framework that integrates Positive-Unlabeled learning with Variational AutoEncoder models for Missing-Not-At-Random imputation and downstream prediction. PU-VAE-MNAR uses PU-learned *indication-driven missingness* information to guide VAE-based MNAR imputation and to improve downstream prediction. Specifically, 1) in the PU-learning step, a predictive adversarial network (PAN) (Hu et al., 2021) estimates the probability that each unlabeled instance belongs to the *positive* set; 2) in the imputation step, the learned probabilities condition the VAE-MNAR model, allowing it to incorporate *indication-driven missingness* information beyond the conventional binary missingness mask; and 3) in the prediction step, the imputed values and PU-learned measurement-intent indicators are jointly used for downstream prediction.

We evaluate PU-VAE-MNAR using simulations, semi-synthetic experiments, and real-world data, assessing both imputation quality and downstream prediction. The semi-synthetic experiments include tabular data with continuous and binary variables under multiple missingness settings. We compare PU-VAE-MNAR with 10 imputation methods across several methodological categories. Across multiple settings, PU-VAE-MNAR achieves superior performance than the evaluated baselines in both tasks.

2 Related Work

Positive-Unlabeled Learning. Positive-Unlabeled (PU) learning addresses binary classification when positive labels are observed but true negatives are unavailable (Bekker & Davis, 2020). Early methods commonly followed a two-step strategy that first identified reliable negatives from the unlabeled set and then trained a classifier using positive, reliable-negative, and remaining unlabeled examples (Lee & Liu, 2003; Li & Liu, 2005). More recent GAN-based approaches generated positive and/or negative examples for subsequent PU classification (Chiaroni et al., 2018; Hou et al., 2018). Many PU methods require estimating the class prior (Xu et al., 2017; Du Plessis et al., 2017; Kiryo et al., 2017; Hou et al., 2018; Chiaroni et al., 2018; Kato et al., 2019), which can be difficult in

high-missing-rate MNAR settings. Predictive Adversarial Networks (PAN) (Hu et al., 2021) avoid this requirement by directly identifying likely positives within the unlabeled set. Few studies have connected PU learning with subsequent missing-value modeling. VAE-PU (Na et al., 2020) uses a VAE to model selection bias and generate virtual PU examples for classification, while Wang et al. (2026) formulates implicit-feedback recommendation as an MNAR PU-learning problem. These methods primarily target label recovery, classification, or debiasing. In contrast, we use PU learning to recover *measurement-intent* information arising under *indication-driven missingness*, and then use this information to guide MNAR imputation and downstream prediction.

Deep Generative Models for MNAR Imputation. Deep generative models have been widely studied for missing-data imputation. GAN-based methods such as GAIN (Yoon et al., 2018) and MisGAN (Li et al., 2019) use adversarial learning to reconstruct missing values, including under some MNAR settings. VAE-based methods more directly model MNAR missingness through latent-variable formulations. not-MIWAE (Ipsen et al., 2021) uses selection models (Little & Rubin, 2019) to jointly model observed values and missingness, GINA (Ma & Zhang, 2021) extends this framework with an identifiable latent-variable model, PSMVAE (Ghalebikesabi et al., 2021) models ignorable and nonignorable missingness through pattern-set mixtures, and GNR (Chen et al., 2023) jointly reconstructs data values and missingness masks. These methods generally represent missingness through a binary mask. Under *indication-driven missingness*, however, the binary mask does not capture whether a variable was intended to be measured. The partially observed *measurement-intent* information can therefore provide complementary information beyond the conventional mask and may be informative for both missing-value imputation and downstream prediction (Agniel et al., 2018; Che et al., 2018; Lipton et al., 2016).

3 Problem Formulation

3.1 Missing Value Imputation for Incomplete Data

We consider missing-value imputation for incomplete tabular data. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathcal{X}^{n \times p}$, where n is the number of instances and p is the number of variables. For instance $i \in \{1, \ldots, n\}$, let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ denote the complete data vector. For each entry x_{ij}, define the binary missingness indicator

$$m_{ij} = \mathbf{1}_{\{x_{ij} \text{ is observed}\}}. \tag{1}$$

where $m_{ij} = 1$ if x_{ij} is observed and $m_{ij} = 0$ otherwise. We denote the observed and missing components of $\mathbf{x}_i$ by $\mathbf{x}_i^{\text{obs}}$ and $\mathbf{x}_i^{\text{mis}}$, respectively, with $\mathbf{x}_i = (\mathbf{x}_i^{\text{obs}}, \mathbf{x}_i^{\text{mis}})$. The imputation task is to estimate the unobserved values $\mathbf{x}_i^{\text{mis}}$ from the available data. Under MNAR, the missingness pattern may depend on the unavailable data and therefore carry information about the missing values. In addition to the conventional binary mask, our setting contains partially observed information associated with *indication-driven missingness*, represented through a *measurement-intent* variable defined in Section 3.2.

3.2 Formulating Measurement Intent as PU Data

When *indication-driven missingness* is present, we introduce a binary *measurement-intent* indicator as

$$S_{ij} = \mathbf{1}_{\{x_{ij} \text{ was intended to be measured}\}}. \tag{2}$$

where $S_{ij} = 1$ denotes positive measurement intent and $S_{ij} = 0$ denotes no intent to measure.

Because S_{ij} is only partially observed, we instead observe a PU label $\widetilde{S}_{ij} \in \{0, 1\}$, where $\widetilde{S}_{ij} = 1$ denotes known positives and $\widetilde{S}_{ij} = 0$ denotes an unlabeled instance; see Table 1. Following standard PU-learning notation (Hu et al., 2021; Teisseyre & Mielniczuk, 2026; Teisseyre et al., 2020), for each variable j, we define $\mathcal{P}_j = \{i : \widetilde{S}_{ij} = 1\}$, and $\mathcal{U}_j = \{i : \widetilde{S}_{ij} = 0\}$. The PU representation therefore provides information beyond m_{ij} and is subsequently integrated into MNAR imputation and downstream prediction.

Table 1: Formal PU representation of *measurement intent* under *indication-driven missingness*. The table relates S_{ij}, $\widetilde{S}_{ij}$, and m_{ij}.

PU class	S_{ij}	$\widetilde{S}_{ij}$	m_{ij}	**Interpretation**
Known positive	1	1	1	Measurement intent is observed and x_{ij} is observed.
Known positive	1	1	0	Measurement intent is observed but x_{ij} is missing.
Unlabeled	$\{0,1\}$	0	0	Measurement intent is unknown and x_{ij} is missing.

4 METHODOLOGY

4.1 PU LEARNING OF MEASUREMENT INTENT

For each variable j, we apply the PAN framework of Hu et al. (2021) to $\mathcal{P}_j$ and $\mathcal{U}_j$. A classifier $C_{\psi_j}(\mathbf{a}_i)$ and discriminator $D_{\omega_j}(\mathbf{a}_i)$ are trained separately for each j using the adversarial objective

$$\begin{aligned}\min_{\psi_j}\max_{\omega_j}\mathcal{L}_{\mathrm{PAN}}^{(j)} = & -\sum_{i\in\mathcal{P}_j\cup\mathcal{U}_j} D_{\mathrm{KL}}\left(P_{ij}^{\mathrm{PU}} \,\|\, D_{ij}\right) \\ & + \lambda_{\mathrm{PAN}}\sum_{i\in\mathcal{U}_j}\left[D_{\mathrm{KL}}\left(D_{ij}\,\|\,C_{ij}\right) - D_{\mathrm{KL}}\left(D_{ij}\,\|\,\widetilde{C}_{ij}\right)\right],\end{aligned} \tag{3}$$

where P_{ij}^{PU} is determined by the observed PU label $\widetilde{S}_{ij}$, $\widetilde{C}_{ij}$ denotes the complementary distribution of C_{ij}, and λ_{PAN} controls the adversarial term.

After training, we retain

$$q_{ij} = C_{\widehat{\psi}_j}(\mathbf{a}_i). \tag{4}$$

the PU-learned probability of positive *measurement intent*. The matrix $\mathbf{Q} = (q_{ij})$ is subsequently integrated into the VAE-MNAR model.

4.2 INTEGRATING PU-LEARNED INFORMATION INTO VAE-MNAR

From the observed PU label $\widetilde{S}_{ij}$ and PU-learned probability q_{ij} in Section 4.1, we define

$$q_{ij}^* = \begin{cases} 1, & \text{if } i \in \mathcal{P}_j, \\ q_{ij}, & \text{if } i \in \mathcal{U}_j. \end{cases} \tag{5}$$

so that known positives are preserved while q_{ij} are used only for unlabeled instances, which avoids introducing additional noise (Zhu et al., 2023b). We then integrate the vector $\mathbf{q}_i^*$ into a VAE-based MNAR model. Following the auxiliary-conditioned prior formulation in GINA, we use $p(\mathbf{z}_i \mid \mathbf{a}_i, \mathbf{q}_i^*)$ and jointly model $\mathbf{x}_i^{\mathrm{obs}}$ and $\mathbf{m}_i$ through the shared latent representation (Chen et al., 2023):

$$p_{\theta,\phi}\left(\mathbf{x}_i^{\mathrm{obs}}, \mathbf{m}_i \mid \mathbf{a}_i, \mathbf{q}_i^*\right) = \int p_\theta\left(\mathbf{x}_i^{\mathrm{obs}} \mid \mathbf{z}_i, \mathbf{a}_i, \mathbf{q}_i^*\right) p_\phi\left(\mathbf{m}_i \mid \mathbf{z}_i, \mathbf{a}_i, \mathbf{q}_i^*\right) p\left(\mathbf{z}_i \mid \mathbf{a}_i, \mathbf{q}_i^*\right) d\mathbf{z}_i. \tag{6}$$

The log-likeliood and the evidence lower bound(ELBO) for sample i is

$$\begin{aligned}\log p_{\theta,\phi}\left(\mathbf{x}_i^{\mathrm{obs}}, \mathbf{m}_i \mid \mathbf{a}_i, \mathbf{q}_i^*\right) &\geq \mathcal{L}_i(\theta,\phi,\gamma) \\ &= \mathbb{E}_{q_\gamma}\left[\log p_\theta\left(\mathbf{x}_i^{\mathrm{obs}} \mid \mathbf{z}_i, \mathbf{a}_i, \mathbf{q}_i^*\right) + \log p_\phi\left(\mathbf{m}_i \mid \mathbf{z}_i, \mathbf{a}_i, \mathbf{q}_i^*\right)\right] \\ &\quad - D_{\mathrm{KL}}\left(q_\gamma \| p\left(\mathbf{z}_i \mid \mathbf{a}_i, \mathbf{q}_i^*\right)\right),\end{aligned} \tag{7}$$

where q_γ denotes the variational posterior (encoder distribution), $q_\gamma\left(\mathbf{z}_i \mid \mathbf{x}_i^{\mathrm{obs}}, \mathbf{m}_i, \mathbf{a}_i, \mathbf{q}_i^*\right)$, and the latent variable is sampled using the standard reparameterization trick.

The standard variational objective is $(\theta^*, \phi^*, \gamma^*) = \arg\max_{\theta,\phi,\gamma}\sum_{i=1}^n \mathcal{L}_i(\theta,\phi,\gamma)$. In PU-VAE-MNAR, we additionally supervise the shared latent model using the PU-learned *measurement-intent* information. Accordingly, the implemented objective is

$$\mathcal{J} = \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_M\mathcal{L}_M + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}} + \lambda_{\mathrm{PU}}\mathcal{L}_{\mathrm{PU}}, \tag{8}$$

where $\mathcal{L}_{\mathrm{PU}}$ is a soft-target binary cross-entropy term that encourages the latent representation to preserve the PU-learned information with some weights. Letting Θ denote all trainable parameters, the implemented optimization problem is therefore $\Theta^* = \arg\min_\Theta \mathcal{J}(\Theta)$. Detailed loss definitions and hyperparameter settings are given in Appendices B and C.4.

Imputation and prediction. We approximate the posterior predictive distribution of $\mathbf{x}_i^{\mathrm{mis}}$ after training as $p\big(\mathbf{x}_i^{\mathrm{mis}} \mid \mathbf{x}_i^{\mathrm{obs}}, \mathbf{m}_i, \mathbf{a}_i, \mathbf{q}_i^*\big) \approx \int p_{\theta^*}\big(\mathbf{x}_i^{\mathrm{mis}} \mid \mathbf{z}_i, \mathbf{a}_i, \mathbf{q}_i^*\big)\, q_{\gamma^*}\big(\mathbf{z}_i \mid \mathbf{x}_i^{\mathrm{obs}}, \mathbf{m}_i, \mathbf{a}_i, \mathbf{q}_i^*\big)\, d\mathbf{z}_i$. Its mean defines the point imputation $\mathbf{x}_i^{\mathrm{imp}}$. The completed vector $\widehat{\mathbf{x}}_i$ retains the observed entries $\mathbf{x}_i^{\mathrm{obs}}$ and replaces $\mathbf{x}_i^{\mathrm{mis}}$ with $\mathbf{x}_i^{\mathrm{imp}}$.

Let $Y_i \in \{0, 1\}$ denote the downstream binary outcome. We use logistic regression as an example of a downstream prediction model. Using $\mathbf{a}_i$ and the completed data $\widehat{\mathbf{x}}_i$, the primary prediction model is

$$\mathrm{logit}\left\{P(Y_i = 1 \mid \mathbf{a}_i, \widehat{\mathbf{x}}_i)\right\} = \beta_0 + \boldsymbol{\beta}_a^\top \mathbf{a}_i + \boldsymbol{\beta}_x^\top \widehat{\mathbf{x}}_i. \tag{9}$$

Other predictor sets are also explored, see Table 4. Importantly, in PU-VAE-MNAR, the outcome labels Y_i are used only to fit the downstream prediction model and are not used to train either the PU-learning model or the VAE-MNAR imputation model (Xie et al., 2020; Le Morvan & Varoquaux, 2025). Therefore, any improvement in downstream prediction using $\widehat{\mathbf{x}}_i$ is attributable to measurement-intent information learned independently of outcome supervision.

5 Experiments

In this section, we conduct simulation, semi-synthetic analysis and real data analysis to study the efficacy of the proposed PU-VAE-MNAR in MNAR imputation and downstream prediction tasks.

5.1 Simulation Study with MNAR Data and PU Labels

Building on the synthetic experimental design of GINA (Ma & Zhang, 2021), we construct three scenarios, denoted A, B, and C, with three continuous variables X_1, X_2, and X_3 generated from independent latent variables $Z_1, Z_2, Z_3 \sim \mathcal{N}(0, 1)$.

The scenarios use different structural equations with additive Gaussian noise; details are given in Appendix C.2. In all scenarios, X_1 is fully observed and used as $\mathbf{a}_i$, whereas X_2 and X_3 are subject to MNAR missingness. Scenario A sets X_j to missing whenever $X_j > 0$, for $j \in \{2, 3\}$. Scenarios B and C use latent-dependent self-masking, with X_j missing whenever $g_j(X_j, Z_1, Z_2, Z_3) > 0$.

To evaluate PU learning with known ground truth, in Scenario A we define the latent measurement-intent indicator as $S_{ij} = \mathbf{1}_{\{X_{ij} \le 1\}}$ for $j \in \{2, 3\}$. The observed PU labels $\widetilde{S}_{ij}$ are generated so that $\mathcal{P}_j \subset \{i : S_{ij} = 1\}$ and the remaining instances form $\mathcal{U}_j$. Thus, $\mathcal{U}_j$ contains unlabeled positives with $0 < X_{ij} \le 1$ and true negatives with $X_{ij} > 1$. The latent S_{ij} is used only for evaluation.

We train PU-VAE-MNAR using both $\mathbf{X}$ and the observed PU labels $\widetilde{\mathbf{S}}$, whereas all baseline models use $\mathbf{X}$ alone. After training, we generate random samples from each model, with additional details provided in Appendix C.2. By comparing the generated samples with the ground-truth data density, we evaluate whether the additional information contained in the PU labels improves recovery of the unavailable data distribution, as shown in Figure 1. PU-VAE-MNAR generates samples that most closely recover the ground-truth distribution after integrating the additional PU-learned information. PVAE(Ma et al., 2019) performs worst and fails to recover the MNAR distribution, while not-MIWAE and GINA show partial recovery but remain biased toward the observed data distribution. These results illustrate the benefit of integrating the simulated *measurement-intent* indicator as additional information for MNAR imputation.

5.2 Real-Data and Semi-Synthetic Experiments

Real-world tabular EHR data. We evaluate PU-VAE-MNAR on a large-scale Cerner electronic health record (EHR) cohort for HIV incidence prediction (Wu et al., 2020; Zhu et al., 2023a; Yu et al., 2020; Brady et al., 2022). The dataset contains 106,261 subjects with binary outcome Y_i indicating incident HIV during the prediction period. The 324 predictors include 157 fully observed covariates $\mathbf{a}_i$ and 167 partially observed laboratory variables $\mathbf{x}_i$, comprising both continuous and

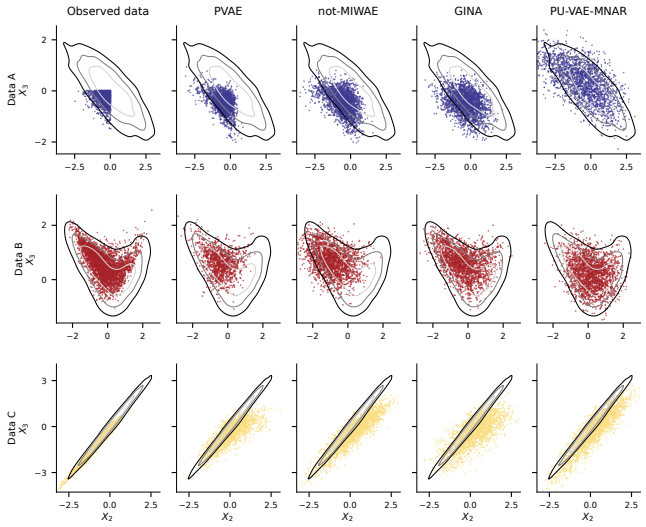


Figure 1: **Generated X_2 and X_3 distributions under synthetic MNAR settings.** Rows correspond to Scenarios A–C. Columns show the observed data, PVAE, not-MIWAE, GINA, and PU-VAE-MNAR, respectively. Points represent observed or model-generated samples, and contour lines indicate the ground-truth density of the complete data.

binary measurements. The laboratory variables have an average missingness rate of 91.9%. This setting provides a natural example of *indication-driven missingness* as shown in Table 1. For each laboratory variable j, an observed indication that the test was ordered or intended to be measured defines the known-positive set $\mathcal{P}_j$. The remaining entries belong to $\mathcal{U}_j$ and may represent either unrecorded measurement intent or true absence of intent. The binary indicator m_{ij} separately records whether x_{ij} is observed. Thus, *measurement intent* and value availability provide distinct information, allowing the PU-learned information to supplement m_{ij}. Laboratory missingness in EHR data can be informative (Agniel et al., 2018; Tan et al., 2023; Li et al., 2021; Sun et al., 2024), because testing decisions depend on patient condition and clinical need. We therefore treat MNAR as a plausible missingness mechanism for these laboratory variables.

Semi-synthetic data. Because the ground-truth values of naturally missing laboratory entries are unavailable, we construct semi-synthetic experiments by additionally masking laboratory values that are originally observed, thereby retaining their ground-truth values for evaluation. We consider artificial masking rates of $r \in \{5\%, 10\%, 30\%, 50\%, 70\%, 90\%, 95\%, 99\%\}$ and evaluate both imputation quality and downstream prediction performance. Additional cohort construction and preprocessing details are provided in Appendix C.3.

Baselines. We compare PU-VAE-MNAR with 10 imputation methods spanning four categories: 1) statistical imputation methods, including Mean, Median, and MICE (Buuren & Groothuis-Oudshoorn, 2011); 2) machine-learning imputation using MissForest (Stekhoven & Bühlmann, 2012); 3) GAN-based deep generative methods, including GAIN (Yoon et al., 2018) and MisGAN (Li et al., 2019); and 4) VAE-based deep generative methods, including not-MIWAE(Ipsen et al., 2021), PSMVAE(Ghalebikesabi et al., 2021), GNR(Chen et al., 2023), and GINA(Ma & Zhang, 2021). All competing methods are evaluated using the same patient splits, artificially masked entries, and downstream prediction protocol. Detailed baseline configurations and hyperparameter settings are provided in Appendix C.5.

Evaluation protocol. We split the cohort into 56% training, 14% validation, and 30% independent test sets, using the same patient-level partition across all methods and masking rates. Models are trained on the training set, the validation set is used for model selection and hyperparameter tuning, and the test set is reserved for final evaluation. Imputation is evaluated only on artificially masked entries with known ground truth. For continuous variables, we report RMSE and pooled RMSE, with pooled RMSE assigning equal weight to each artificially masked entry, after min-max normalization using training set parameters. For binary variables, we report F1 at a fixed threshold of 0.5 and the

Table 2: MNAR imputation performance. Lower is better for RMSE, pooled RMSE, and Brier score (↓), while higher is better for F1 (↑). **Bold** indicates the best result in each column.

Method	Overall		Continuous variables		Binary variables	
	RMSE ↓	Pooled RMSE ↓	RMSE ↓	Pooled RMSE ↓	Brier ↓	F1 ↑
Mean	0.127	0.138	0.067	0.066	0.113	0.325
Median	0.138	0.155	0.069	0.067	0.149	0.325
MICE	0.114	0.122	0.056	0.053	0.093	0.523
MissForest	0.122	0.139	0.055	0.050	0.127	0.554
MisGAN	0.270	0.318	0.235	0.289	0.215	0.457
GAIN	0.271	0.257	0.208	0.208	0.212	0.315
GINA	0.123	0.133	0.064	0.063	0.106	0.396
not-MIWAE	0.156	0.146	0.120	0.099	0.095	0.496
PSMVAE	0.155	0.177	0.074	0.071	0.198	0.277
GNR	0.118	0.128	0.065	0.064	0.096	0.489
PU-VAE-MNAR	**0.110**	**0.120**	**0.053**	**0.048**	**0.088**	**0.590**

Table 3: Downstream prediction performance. Higher is better for all metrics. Sens. = sensitivity, Spec. = specificity, NPV = negative predictive value, and PPV = positive predictive value. **Bold** indicates the best result in each column.

Method	AUC	AUCPR	R^2	Sens.	Spec.	NPV	PPV
Mean	0.830	0.718	0.409	0.671	0.841	0.841	0.661
Median	0.833	0.724	0.417	0.686	0.834	0.849	0.650
MICE	0.877	0.753	0.443	0.703	0.892	0.890	0.708
MissForest	0.912	0.807	0.514	0.715	0.936	0.910	0.784
MisGAN	0.911	0.812	0.513	0.756	0.918	0.920	0.750
GAIN	0.911	0.812	0.514	0.730	0.932	0.914	0.776
GINA	0.914	0.792	0.477	0.763	0.916	0.934	0.720
not-MIWAE	0.919	0.830	0.538	0.766	0.924	0.924	0.765
PSMVAE	0.899	0.778	0.472	0.743	0.906	0.915	0.720
GNR	0.919	0.823	0.530	0.756	0.926	0.921	0.770
PU-VAE-MNAR	**0.936**	**0.864**	**0.603**	**0.804**	**0.938**	**0.951**	**0.807**

Brier score. Downstream prediction is evaluated on the held-out test set after imputation. For logistic regression, the penalty is selected using the training data and the decision threshold using the validation set. Prediction metrics include AUC, AUCPR, R^2, sensitivity, specificity, NPV, and PPV. Additional dataset, model, and implementation details are provided in Appendix C.3.

5.2.1 Results: Real-World MNAR Imputation and Prediction

We first evaluate PU-VAE-MNAR MNAR imputation of the partially observed laboratory variables. For categorical laboratory features, we use binary indicators (e.g., normal versus abnormal). Table 2 compares PU-VAE-MNAR with the baseline methods for both continuous and binary variables. We then evaluate downstream HIV incidence prediction using $\widehat{\mathbf{x}}_i$ together with $\mathbf{a}_i$. Table 3 compares the predictive performance of PU-VAE-MNAR and the competing imputation methods.

Strong performance of PU-VAE-MNAR. PU-VAE-MNAR achieves the best overall imputation performance in Table 2 across both continuous and binary variables, outperforming selected baselines. This suggests that the PU-learned *measurement-intent* information provides complementary information beyond m_{ij}. The advantage extends to downstream prediction. As shown in Table 3, PU-VAE-MNAR achieves the best overall predictive performance. Because Y_i is not used to train either the PU or VAE-MNAR model, these gains arise from information learned independently of

Table 4: Prediction performance with fixed imputer and varied predictors. **Bold** indicates the best reported value in each column. Higher values are better for all reported metrics. $\mathbf{u}$ denotes the deterministic uncertainty score defined in Section 5.3; Sens., sensitivity; Spec., specificity; NPV, negative predictive value; PPV, positive predictive value. †Input denotes the input predictor dimension.

Predictor set	Input†	AUC	AUCPR	R^2	Sens.	Spec.	NPV	PPV
$\mathbf{a}, \widetilde{\mathbf{S}}$	324	0.921	0.835	0.561	0.793	0.921	0.935	0.772
$\mathbf{a}, \hat{\mathbf{x}}$ (PU-VAE-MNAR)	324	0.936	0.864	0.603	0.804	0.938	0.951	0.807
$\mathbf{a}, \mathbf{q}^*$	324	0.945	0.889	0.650	0.817	0.950	0.941	0.841
$\mathbf{a}, \hat{\mathbf{x}}, \mathbf{m}$	491	0.939	0.871	0.633	0.812	0.939	0.938	0.823
$\mathbf{a}, \hat{\mathbf{x}}, \mathbf{q}^*$	491	0.949	0.902	0.675	0.812	**0.961**	0.940	**0.870**
$\mathbf{a}, \hat{\mathbf{x}}, \mathbf{q}^*, \mathbf{u}$	658	0.958	0.925	0.720	0.878	0.942	0.960	0.831
$\mathbf{a}, \hat{\mathbf{x}}, \widetilde{\mathbf{S}}, \mathbf{q}, \mathbf{u}$	825	**0.962**	**0.932**	**0.737**	**0.886**	0.942	**0.962**	0.832

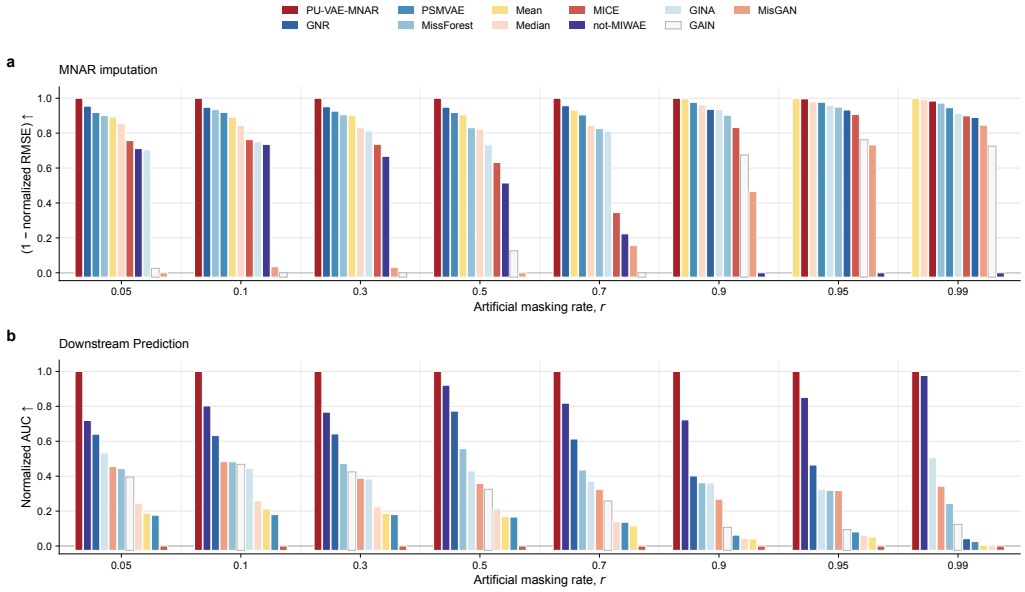


Figure 2: Semi-synthetic evaluation across artificial masking rates. The horizontal axis denotes the artificial masking rate r. (a) MNAR imputation performance, shown as $1 - \text{normalized RMSE}$. (b) Downstream prediction performance, shown as normalized AUC. Metrics are normalized within each masking rate for visualization, and higher values indicate better performance in both panels.

outcome supervision. Together, the results show improved MNAR imputation and greater downstream utility of the completed data with integrated PU-learned information.

Different algorithmic preferences in tabular data. The competing methods show different behavior across continuous and binary variables. MissForest and several VAE-based methods perform relatively well for continuous variables but are less competitive for binary variables, consistent with prior evidence that heterogeneous tabular data may favor different modeling strategies (Zhang et al., 2025; El Badisy et al., 2024; Ma et al., 2020). Among other VAEs, selection-model approaches such as not-MIWAE differ from pattern-set mixture methods such as PSMVAE, which may be less suited to highly sparse, feature-specific missingness patterns. GNR performs relatively close to PU-VAE-MNAR, but PU-VAE-MNAR consistently achieves better overall performance by integrating additional PU-learned *measurement-intent* information beyond the binary missingness mask.

Table 5: Prediction performance with a fixed predictor set under different imputer configurations. Input denotes the downstream predictor dimension. VAE-MNAR excludes all PU-derived inputs and objectives. PUU-VAE-MNAR additionally conditions the imputer on the uncertainty score $\mathbf{u}$ together with $\mathbf{q}^*$. Higher is better for all metrics, and **bold** indicates the best result in each column.

Imputer	Input	AUC	AUCPR	R^2	Sens.	Spec.	NPV	PPV
PU-VAE-MNAR	324	**0.936**	**0.864**	**0.603**	**0.804**	**0.938**	**0.951**	**0.807**
VAE-MNAR	324	0.913	0.816	0.524	0.746	0.927	0.918	0.769
PUU-VAE-MNAR	324	0.922	0.832	0.550	0.796	0.916	0.932	0.756

5.2.2 Results: Semi-Synthetic Experiments

Impact of PU learning in heterogeneous tabular data. As shown in Figure 2, PU-VAE-MNAR achieves the lowest RMSE at lower masking rates, although imputation performance becomes less stable under extreme sparsity and can fall below simple mean or median imputation. In contrast, PU-VAE-MNAR consistently achieves the best downstream prediction across masking rates, indicating that the additional PU-learned *measurement-intent* information remains useful even when accurate imputation becomes difficult (Sisk et al., 2021; Sun et al., 2024). GAIN also shows that imputation and prediction need not improve together: despite relatively poor imputation in some settings, it retains competitive predictive performance, consistent with prior findings (Le Morvan & Varoquaux, 2025). PU-VAE-MNAR further improves downstream prediction by integrating information associated with *indication-driven missingness* beyond the conventional missingness mask.

5.3 Ablation Studies

To disentangle how PU-learned information contributes to downstream prediction, we perform two ablation analyses. First, we fix the imputer and vary the prediction inputs among $\mathbf{x}_i^{\text{imp}}$, $\mathbf{m}_i$, $\mathbf{q}_i^*$, and the deterministic uncertainty score $u_{ij} = (1 - \widetilde{S}_{ij})q_{ij}^*(1 - q_{ij}^*)$. Second, we fix the downstream predictor and vary the information provided to the imputer to assess whether $\mathbf{q}_i^*$ improves prediction through better imputation or as an additional predictive feature. Input dimensions are also reported in Table 4. With the imputer fixed, adding $\mathbf{m}_i$, $\mathbf{q}_i^*$, and $(\mathbf{q}_i^*, \mathbf{u}_i)$ increases AUC. Because $\mathbf{m}_i$ and $\mathbf{q}_i^*$ have the same dimensionality, the improvement from $\mathbf{q}_i^*$ reflects additional information rather than a larger predictor set. When varying the imputer, incorporating $\mathbf{q}_i^*$ also outperforms the $(\widehat{\mathbf{x}}_i, \mathbf{m}_i)$ configuration, indicating that PU-learned *measurement-intent* information improves the imputed representation itself. In contrast, $\mathbf{u}_i$ is more useful as a downstream predictive feature than as an imputation input, as shown in Table 5.

6 Conclusions

In this paper, we introduced *indication-driven missingness* as a distinct missing-data setting in which measurement is selectively initiated by subject-specific indications, and formulated the resulting partially observed *measurement intent* as a PU-learning problem. Building on this formulation, we proposed PU-VAE-MNAR, which incorporates PU-learned information beyond the conventional binary missingness mask into VAE-based MNAR imputation and downstream prediction. Across simulation, semi-synthetic, and real-world experiments, PU-VAE-MNAR improves distribution recovery, imputation quality, and predictive performance across a range of missingness settings. Ablation studies further show that information associated with *indication-driven missingness* can benefit prediction both indirectly through improved imputation and directly as complementary predictive information beyond the imputed values. These results highlight the value of explicitly modeling *indication-driven missingness* when measurement intent is only partially observed.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing, for retrieval and discovery of related work, and for research ideation or execution. We have not used generative AI tools to draft sections of the paper, generate synthetic datasets, or prove mathematical claims. We have reviewed all AI-assisted work. In particular, AI-assisted writing was manually reviewed and revised by the authors, retrieved literature and references were checked against the original sources, and AI-assisted research ideation were independently evaluated by the authors. We take responsibility for the final content of this work, including text, claims, references, analyses, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of the data preprocessing procedures, model architecture, training procedure, hyperparameter settings, and experimental protocols in the main text and appendices. All model development and evaluation follow predefined train, validation, and test splits to avoid information leakage. The Cerner Health Facts® data cannot be publicly redistributed because of data-use and privacy restrictions. The proposed methodology, implementation details, and experiments on synthetic data are described to enable reproduction of the methodological results independent of access to the proprietary EHR data.

REFERENCES

Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: Retrospective observational study. *BMJ*, pp. k1479, April 2018. ISSN 0959-8138, 1756-1833. doi: 10.1136/bmj.k1479.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, April 2020. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-020-05877-5.

Veronica Brady, Meagan Whisenant, Xueying Wang, Vi K. Ly, Gen Zhu, David Aguilar, and Hulin Wu. Characterization of symptoms and symptom clusters for type 2 diabetes using a large nationwide electronic health record database. *Diabetes Spectrum*, 35(2):159–170, January 2022. ISSN 1040-9165. doi: 10.2337/ds21-0064.

Stef Van Buuren and Karin Groothuis-Oudshoorn. **Mice** : Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, 45(3), 2011. ISSN 1548-7660. doi: 10.18637/jss.v045.i03.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1): 6085, April 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-24271-9.

Jialei Chen, Yuanbo Xu, Pengyang Wang, and Yongjian Yang. Deep Generative Imputation Model for Missing Not At Random Data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 316–325, October 2023. doi: 10.1145/3583780.3614835.

F. Chiaroni, M-C. Rahal, N. Hueber, and F. Dufaux. Learning with A Generative Adversarial Network From a Positive Unlabeled Dataset for Image Classification. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 1368–1372, Athens, October 2018. IEEE. ISBN 978-1-4799-7061-2. doi: 10.1109/ICIP.2018.8451831.

Zongyu Dai, Zhiqi Bu, and Qi Long. Multiple Imputation via Generative Adversarial Network for High-dimensional Blockwise Missing Value Problems. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 791–798, Pasadena, CA, USA, December 2021. IEEE. ISBN 978-1-6654-4337-1. doi: 10.1109/ICMLA52953.2021.00131.

Marthinus C. Du Plessis, Gang Niu, and Masashi Sugiyama. Class-prior estimation for learning from positive and unlabeled data. *Machine Learning*, 106(4):463–492, April 2017. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-016-5604-6.

Imad El Badisy, Nathalie Graffeo, Mohamed Khalis, and Roch Giorgi. Multi-metric comparison of machine learning imputation methods with application to breast cancer survival. *BMC Medical Research Methodology*, 24(1):191, August 2024. ISSN 1471-2288. doi: 10.1186/s12874-024-02305-3.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 213–220, Las Vegas Nevada USA, August 2008. ACM. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401920.

Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. A survey on missing data in machine learning. *Journal of Big Data*, 8(1):140, October 2021. ISSN 2196-1115. doi: 10.1186/s40537-021-00516-9.

Sahra Ghalebikesabi, Rob Cornish, Chris Holmes, and Luke Kelly. Deep generative missingness pattern-set mixture models. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3727–3735. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/ghalebikesabi21a.html.

Ming Hou, Brahim Chaib-draa, Chao Li, and Qibin Zhao. Generative Adversarial Positive-Unlabelled Learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 2255–2261, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-2-7. doi: 10.24963/ijcai.2018/312.

Wenpeng Hu, Ran Le, Bing Liu, Feng Ji, Jinwen Ma, Dongyan Zhao, and Rui Yan. Predictive Adversarial Learning from Positive and Unlabeled Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):7806–7814, May 2021. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v35i9.16953.

Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. not-{miwae}: Deep generative modelling with missing not at random data. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tu29GQT0JFy.

Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational autoencoder with arbitrary conditioning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxtJh0qYm.

Kai Jiang and Tru Cao. Automated hiv case identification from the mimic-iv database. *AMIA Jt Summits Transl Sci Proc*, 2024:555–564, 2024. ISSN 2153-4063 (Electronic).

Masahiro Kato, Takeshi Teshima, and Junya Honda. Learning from positive and unlabeled data with a selection bias. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJzLciCqKm.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/7cce53cf90577442771720a370c3c723-Paper.pdf.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 275–284, Halifax NS Canada, August 2017. ACM. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098066.

Marine Le Morvan and Gaël Varoquaux. Imputation for prediction: Beware of diminishing returns. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/65e837e76a5308df3d5544aab6196e21-Abstract-Conference.html.

Wee Sun Lee and Bing Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pp. 448–455. AAAI Press, 2003. ISBN 1577351894.

Jiang Li, Xiaowei S. Yan, Durgesh Chaudhary, Venkatesh Avula, Satish Mudiganti, Hannah Husby, Shima Shahjouei, Ardavan Afshar, Walter F. Stewart, Mohammed Yeasin, Ramin Zand, and Vida Abedi. Imputation of missing values for electronic health record laboratory data. *npj Digital Medicine*, 4(1):147, 2021. doi: 10.1038/s41746-021-00518-0. URL https://doi.org/10.1038/s41746-021-00518-0.

Steven Cheng-Xian Li, Bo Jiang, and Benjamin M. Marlin. MisGAN: Learning from incomplete data with generative adversarial networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1lDV3RcKm.

Xiao-Li Li and Bing Liu. Learning from Positive and Unlabeled Examples with Different Data Distributions. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, João Gama, Rui Camacho, Pavel B. Brazdil, Alípio Mário Jorge, and Luís Torgo (eds.), *Machine Learning: ECML 2005*, volume 3720, pp. 218–229. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005. ISBN 978-3-540-29243-2 978-3-540-31692-3. doi: 10.1007/11564096_24.

Zachary C Lipton, David Kale, and Randall Wetzel. Directly modeling missing data in sequences with rnns: Improved classification of clinical time series. In Finale Doshi-Velez, Jim Fackler, David Kale, Byron Wallace, and Jenna Wiens (eds.), *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pp. 253–270, Northeastern University, Boston, MA, USA, 18–19 Aug 2016. PMLR. URL https://proceedings.mlr.press/v56/Lipton16.html.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, 3 edition, 2019. ISBN 9780470526798. doi: 10.1002/9781119482260. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781119482260.

Todd D Little and Mijke Rhemtulla. Planned Missing Data Designs for Developmental Researchers. *Child Development Perspectives*, 7(4):199–204, December 2013. ISSN 1750-8592, 1750-8606. doi: 10.1111/cdep.12043.

Chao Ma and Cheng Zhang. Identifiable generative models for missing not at random data imputation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=bGXIX-CVzrq.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient dynamic discovery of high-value information with partial VAE. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4234–4243. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ma19c.html.

Chao Ma, Sebastian Tschiatschek, Richard E. Turner, José Miguel Hernández-Lobato, and Cheng Zhang. VAEM: a deep generative model for heterogeneous mixed type data. In *ICML Workshop on the Art of Learning with Missing Values (Artemiss)*, 2020. URL https://openreview.net/forum?id=JZ-6j-siNBj.

Zhihua Ma and Guanghui Chen. Bayesian methods for dealing with missing data problems. *Journal of the Korean Statistical Society*, 47(3):297–313, September 2018. ISSN 12263192. doi: 10.1016/j.jkss.2018.03.002.

Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: Deep generative modelling and imputation of incomplete data sets. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4413–4423. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/mattei19a.html.

Della Murbarani Prawidya Murti, Utomo Pujianto, Aji Prasetya Wibawa, and Muhammad Iqbal Akbar. K-Nearest Neighbor (K-NN) based Missing Data Imputation. In *2019 5th International Conference on Science in Information Technology (ICSITech)*, pp. 83–88, Yogyakarta, Indonesia, October 2019. IEEE. ISBN 978-1-7281-2378-3 978-1-7281-2380-6. doi: 10.1109/ICSITech46713.2019.8987530.

Byeonghu Na, Hyemi Kim, Kyungwoo Song, Weonyoung Joo, Yoon-Yeong Kim, and Il-Chul Moon. Deep Generative Positive-Unlabeled Learning under Selection Bias. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 1155–1164, Virtual Event Ireland, October 2020. ACM. ISBN 978-1-4503-6859-9. doi: 10.1145/3340531.3411971.

Swj Nijman, Am Leeuwenberg, I Beekers, I Verkouter, Jjl Jacobs, Ml Bots, Fw Asselbergs, Kgm Moons, and Tpa Debray. Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review. *Journal of Clinical Epidemiology*, 142:218–229, February 2022. ISSN 08954356. doi: 10.1016/j.jclinepi.2021.11.023.

Filip Radlinski and Thorsten Joachims. Query chains: Learning to rank from implicit feedback. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 239–248, Chicago Illinois USA, August 2005. ACM. ISBN 978-1-59593-135-1. doi: 10.1145/1081870.1081899.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2335739.

Joseph L. Schafer and John W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002. ISSN 1939-1463, 1082-989X. doi: 10.1037/1082-989X.7.2.147.

Rose Sisk, Lijing Lin, Matthew Sperrin, Jessica K Barrett, Brian Tom, Karla Diaz-Ordaz, Niels Peek, and Glen P Martin. Informative presence and observation in routine health data: A review of methodology for clinical risk prediction. *Journal of the American Medical Informatics Association*, 28(1):155–166, 01 2021. ISSN 1527-974X. doi: 10.1093/jamia/ocaa242. URL https://doi.org/10.1093/jamia/ocaa242.

Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, January 2012. ISSN 1367-4811, 1367-4803. doi: 10.1093/bioinformatics/btr597.

Minghui Sun, Matthew M. Engelhard, Armando D. Bedoya, and Benjamin A. Goldstein. Incorporating informatively collected laboratory data from ehr in clinical prediction models. *BMC Medical Informatics and Decision Making*, 24(1):206, 2024. doi: 10.1186/s12911-024-02612-1. URL https://doi.org/10.1186/s12911-024-02612-1.

Amelia L.M. Tan, Emily J. Getzen, Meghan R. Hutch, Zachary H. Strasser, Alba Gutiérrez-Sacristán, Trang T. Le, Arianna Dagliati, Michele Morris, David A. Hanauer, Bertrand Moal, Clara-Lea Bonzel, William Yuan, Lorenzo Chiudinelli, Priam Das, Harrison G. Zhang, Bruce J. Aronow, Paul Avillach, Gabriel.A. Brat, Tianxi Cai, Chuan Hong, William G. La Cava, He Hooi Will Loh, Yuan Luo, Shawn N. Murphy, Kee Yuan Hgiam, Gilbert S. Omenn, Lav P. Patel, Malarkodi Jebathilagam Samayamuthu, Emily R. Shriver, Zahra Shakeri Hossein Abad, Byorn W.L. Tan, Shyam Visweswaran, Xuan Wang, Griffin M. Weber, Zongqi Xia, Bertrand Verdy, Qi Long, Danielle L. Mowery, and John H. Holmes. Informative missingness: What can we learn from patterns in missing laboratory data in the electronic health record? *Journal of Biomedical Informatics*, 139:104306, March 2023. ISSN 15320464. doi: 10.1016/j.jbi.2023.104306.

Paweł Teisseyre and Jan Mielniczuk. Positive-unlabeled regression: learning from partially labeled quantitative outcomes. In *Forty-Second Annual Conference on Uncertainty in Artificial Intelligence*, 2026. URL https://openreview.net/forum?id=37CIw1arIb.

Paweł Teisseyre, Jan Mielniczuk, and Małgorzata Łazęcka. Different strategies of fitting logistic regression for positive and unlabelled data. In Valeria V. Krzhizhanovskaya, Gábor Závodszky, Michael H. Lees, Jack J. Dongarra, Peter M. A. Sloot, Sérgio Brissos, and João Teixeira (eds.), *Computational Science – ICCS 2020*, volume 12140, pp. 3–17. Springer International Publishing, Cham, 2020. doi: 10.1007/978-3-030-50423-6_1.

Sichao Wang, Tianyu Xia, and Lingxiao Yang. Positive-unlabeled learning in implicit feedback from data missing-not-at-random perspective. *Entropy*, 28(1):41, 2026. doi: 10.3390/e28010041. URL https://www.mdpi.com/1099-4300/28/1/41.

Hulin Wu, Jose Miguel Yamal, Ashraf Yaseen, and Vahed Maroufy. *Statistics and Machine Learning Methods for EHR Data: From Data Extraction to Data Analytics*. Chapman & Hall/CRC Healthcare Informatics Series. CRC Press, Boca Raton, FL, 1 edition, 2020. doi: 10.1201/9781003030003.

Ruimin Xie, Nabil Magbool Jan, Kuangrong Hao, Lei Chen, and Biao Huang. Supervised Variational Autoencoders for Soft Sensor Modeling With Missing Data. *IEEE Transactions on Industrial Informatics*, 16(4):2820–2828, April 2020. ISSN 1551-3203, 1941-0050. doi: 10.1109/TII.2019.2951622.

Yixing Xu, Chang Xu, Chao Xu, and Dacheng Tao. Multi-positive and unlabeled learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 3182–3188, 2017. doi: 10.24963/ijcai.2017/444. URL https://doi.org/10.24963/ijcai.2017/444.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5689–5698. PMLR, 2018. URL https://proceedings.mlr.press/v80/yoon18a.html.

Duo Yu, George W Williams, David Aguilar, José-Miguel Yamal, Vahed Maroufy, Xueying Wang, Chenguang Zhang, Yuefan Huang, Yuxuan Gu, Yashar Talebi, and Hulin Wu. Machine learning prediction of the adverse outcome for nontraumatic subarachnoid hemorrhage patients. *Ann Clin Transl Neurol*, 7(11):2178–2185, Nov 2020. ISSN 2328-9503 (Electronic); 2328-9503 (Linking). doi: 10.1002/acn3.51208.

Hengrui Zhang, Liancheng Fang, Qitian Wu, and Philip S. Yu. Diffputer: Empowering diffusion models for missing data imputation. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=3fl1SENSYO.

Gen Zhu, Anthony Yuan, Duo Yu, Alicia Zha, and Hulin Wu. Machine learning to predict mortality for aneurysmal subarachnoid hemorrhage (asah) using a large nationwide ehr database. *PLOS Digital Health*, 2(12):1–10, 12 2023a. doi: 10.1371/journal.pdig.0000400. URL https://doi.org/10.1371/journal.pdig.0000400.

Zhangchi Zhu, Lu Wang, Pu Zhao, Chao Du, Wei Zhang, Hang Dong, Bo Qiao, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Robust positive-unlabeled learning via noise negative sample self-correction. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 3663–3673, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599491. URL https://doi.org/10.1145/3580305.3599491.

A Definition of symbols and notations

Table 6 summarizes the main notation used in this paper.

Table 6: Definitions of the main symbols.

Symbol	Explanation
$n,\ p;\ i,\ j$	Numbers of instances and variables; corresponding indices.
$\mathbf{X},\ \mathbf{x}_i,\ x_{ij}$	Complete data matrix, instance vector, and entry.
$\mathbf{x}_i^{\text{obs}},\ \mathbf{x}_i^{\text{mis}}$	Observed and missing subvectors.
$\mathbf{m}_i,\ m_{ij}$	Observation mask: 1 = observed, 0 = missing.
$\mathbf{a}_i$	Fully observed auxiliary covariates.
S_{ij}	Measurement intent: 1 = positive, 0 = negative.
$\widetilde{\mathbf{S}},\ \widetilde{S}_{ij}$	PU labels: 1 = known positive, 0 = unlabeled.
$\mathcal{P}_j,\ \mathcal{U}_j$	Known-positive and unlabeled index sets.
$C_{\psi_j};\ C_{ij},\ \widetilde{C}_{ij}$	Classifier; output distribution and its complement.
$D_{\omega_j};\ D_{ij}$	Discriminator; output distribution.
$\psi_j,\ \omega_j;\ \widehat{\psi}_j$	Classifier and discriminator parameters; fitted classifier parameters.
P_{ij}^{PU}	Target distribution defined by the observed PU label.
$\mathcal{L}_{\text{PAN}}^{(j)},\ \lambda_{\text{PAN}}$	PAN objective and adversarial weight.
$\mathbf{Q},\ q_{ij};\ q_{ij}^*,\ \mathbf{q}_i^*$	Raw PU scores; corrected scores with known positives set to 1.
$u_{ij},\ \mathbf{u}_i$	PU uncertainty: $(1-\widetilde{S}_{ij})q_{ij}^*(1-q_{ij}^*)$.
$\mathbf{z}_i$	Shared latent vector.
$\theta,\ \phi,\ \gamma$	Data-decoder, mask-decoder, and encoder parameters.
$p_\theta,\ p_\phi,\ p_{\theta,\phi}$	Data, mask, and joint conditional distributions.
$p(\mathbf{z}_i \mid \mathbf{a}_i, \mathbf{q}_i^*),\ q_\gamma$	Conditional latent prior and variational posterior.
$\mathcal{L}_i$	Per-instance ELBO.
$\mathcal{L}_{\text{cont}},\ \mathcal{L}_{\text{bin}}$	Continuous and binary reconstruction losses.
$\mathcal{L}_{\text{rec}},\ \mathcal{L}_M,\ \mathcal{L}_{\text{KL}}$	Data reconstruction, mask reconstruction, and latent KL losses.
$\widehat{\mathbf{q}}_i,\ \widehat{q}_{ij}$	Auxiliary PU predictions.
$\mathcal{L}_{\text{PU}},\ N_q$	Auxiliary PU loss and number of included entries.
$\lambda_{\text{rec}},\ \lambda_M,\ \lambda_{\text{KL}},\ \lambda_{\text{PU}}$	Weights in the compact VAE objective.
$\lambda_{\text{cont}},\ \lambda_{\text{bin}}$	Weights of continuous and binary reconstruction.
$\mathcal{J};\ \Theta,\ \Theta^*$	Total VAE loss; all trainable parameters and fitted values.
$\theta^*,\ \phi^*,\ \gamma^*$	Fitted decoder and encoder parameters.
$\mathbf{x}_i^{\text{imp}},\ \widehat{\mathbf{x}}_i$	Imputed missing subvector and completed data vector.
Y_i	Binary downstream outcome.
$\beta_0;\ \boldsymbol{\beta}_a,\ \boldsymbol{\beta}_x$	Logistic intercept; coefficients for $\mathbf{a}_i$ and $\widehat{\mathbf{x}}_i$.
$X_j,\ X_{ij};\ Z_j,\ \epsilon_j$	Simulated covariates; latent factors and noise.
$M_j,\ S_j,\ \widetilde{S}_j$	Simulation observation, intent, and PU indicators.
$g_j,\ g;\ b;\ r$	Masking function; outcome intercept; masking rate.

B Additional Model Details

In implementation, the reconstruction term in Eq. 8 is further separated according to variable type, yielding the expanded objective in Eq. 10. For continuous variables, we use a Gaussian negative log-likelihood, whereas binary variables are modeled using a Bernoulli likelihood with binary cross-entropy loss. We therefore write

$$\mathcal{J} = \lambda_{\text{cont}}\mathcal{L}_{\text{cont}} + \lambda_{\text{bin}}\mathcal{L}_{\text{bin}} + \lambda_M\mathcal{L}_M + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{PU}}\mathcal{L}_{\text{PU}}. \tag{10}$$

where $\mathcal{L}_{\text{cont}}$ is the Gaussian negative log-likelihood for continuous variables, $\mathcal{L}_{\text{bin}}$ is the Bernoulli binary cross-entropy loss for binary variables, $\mathcal{L}_M$ is the standard binary cross-entropy loss for reconstruction of the missingness indicator, and $\mathcal{L}_{\text{KL}}$ is the standard KL-divergence regularization

term between the variational posterior and the conditional prior. The PU-learned measurement-intent scores $\mathbf{q}_i^*$ are integrated into VAE training through an auxiliary PU loss objective. Because the reconstruction, missingness, KL, and PU loss terms can have different numerical scales, we assign separate weights to the individual components of the training objective rather than adjusting only the PU loss term.

The shared latent representation $\mathbf{z}_i$ is used to estimate the measurement-intent score vector $\widehat{\mathbf{q}}_i$. The target $\mathbf{q}_i^*$ is fixed during VAE training. The auxiliary contribution is defined using a soft-target binary cross-entropy loss:

$$\lambda_{\mathrm{PU}}\mathcal{L}_{\mathrm{PU}} = -\frac{\lambda_{\mathrm{PU}}}{N_q}\sum_{i,j}\left[q_{ij}^*\log\widehat{q}_{ij} + (1-q_{ij}^*)\log(1-\widehat{q}_{ij})\right], \tag{11}$$

where N_q denotes the number of entries included in the auxiliary loss.

The corresponding loss weights, search ranges, and additional optimization settings are provided in Appendix C.4.

C Experimental Details

C.1 Configurations.

We conduct all experiments with:

- Operating System: Red Hat Enterprise Linux 10.1 (Coughlan)
- CPU: Intel Xeon Gold 6548Y+
- GPU: NVIDIA H100 NVL with 96 GB of Memory
- Software: CUDA 12.1, Python 3.12.11, PyTorch 2.5.1, and R 4.4.2

C.2 Synthetic data generation

We constructed three synthetic datasets, A, B, and C, with three latent variables, three covariates, and a binary outcome. We independently sampled $Z_j \sim \mathcal{N}(0,1)$ and $\epsilon_j \sim \mathcal{N}(0, 0.1^2)$ for $j = 1, 2, 3$, and generated

$$X_1 = 0.8Z_1 + 0.4Z_2 + 0.2Z_3 + \epsilon_1. \tag{12}$$

The other two covariates were generated as follows:

$$\begin{aligned}
\text{A:}\quad & X_2 = 0.8X_1 + 0.6Z_2 + 0.2Z_3 + \epsilon_2,\\
& X_3 = -0.8X_2 + 0.2X_1 + 0.5Z_3 + 0.15\sin(Z_1) + \epsilon_3;\\
\text{B:}\quad & X_2 = 0.5X_1 + \sin(1.3Z_2) + 0.3Z_3 + \epsilon_2,\\
& X_3 = 0.35X_2^2 - 0.6X_2 + 0.3X_1 + 0.45Z_3 + \epsilon_3;\\
\text{C:}\quad & X_2 = 0.7X_1 + 0.5Z_2 + 0.2\tanh(Z_3) + \epsilon_2,\\
& X_3 = 1.2X_2 + 0.15\sin(X_1) + 0.15Z_3 + \epsilon_3.
\end{aligned} \tag{13}$$

Missingness and PU labels. X_1 was fully observed, with $M_1 = S_1 = \widetilde{S}_1 = 1$. For $j = 2, 3$, numerical observation and true intent were defined by

$$\begin{aligned}
\text{A:}\quad & M_j = \mathbf{1}\{X_j \le 0\}, && S_j = \mathbf{1}\{X_j \le 1\},\\
\text{B/C:}\quad & M_j = \mathbf{1}\{g(X_j, Z_1, Z_2, Z_3) \le 0\}, && S_j = \mathbf{1}\{g(X_j, Z_1, Z_2, Z_3) \le 0.5\,\mathrm{SD}(g)\}.
\end{aligned} \tag{14}$$

Here, $M_j = 1$ denotes an observed value. For each partially observed covariate, g is a fixed linear combination of its four arguments, with coefficients sampled independently from $\mathrm{Uniform}(-1, 1)$. The standard deviation of each g was estimated from a separate calibration sample of 100,000 individuals. Observed PU labels were drawn independently conditional on the complete data, with

$$\Pr(\widetilde{S}_j = 1 \mid X_1, X_2, X_3, Z_1, Z_2, Z_3) = M_j + 0.9(1 - M_j)S_j. \tag{15}$$

Thus, all observed entries were labeled positive; missing true positives were labeled with probability 0.9, and true negatives remained unlabeled. Models received $\widetilde{S}_j$ when allowed; S_j was reserved for evaluation.

Outcome and evaluation. For outcome generation, the covariates in the following expression were standardized using complete-data means and standard deviations from the calibration sample:

$$\Pr(Y = 1 \mid X_1, X_2, X_3) = \frac{1}{1 + \exp[-(b + 0.4X_1 + 0.9X_2 - 0.8X_3)]}. \tag{16}$$

The intercept b was calibrated separately for A/B/C to give an average outcome probability of 0.3. Imputers used only observed training values for input standardization and did not use Y. All methods shared the same partitions and masks within each replication. Imputation was evaluated on naturally missing test entries using RMSE; logistic regression on completed covariates was evaluated using AUROC. Generated samples were evaluated separately from imputations.

Training individuals	10,000	Validation individuals	1000
Test individuals	2,000	Independent data replications	20
Generated samples per method	2,000	Complete pairs for KDE contours	2,000

C.3 Real-world Datasets

We use a tabular cohort from Cerner Health Facts, a nationwide de-identified EHR database. The cohort contains 106,261 adults, including 26,089 HIV-1 cases ($Y_i = 1$) and 80,172 clean controls ($Y_i = 0$). Cohort membership is determined using an EHR-based phenotyping algorithm based on diagnoses, medications, and HIV tests (Jiang & Cao, 2024). Predictors are collected during a 12-month observation window preceding a 12-month prediction window.

The predictor set contains 324 variables, including 167 laboratory variables (119 numerical (continuous) and 48 binary) and 157 fully observed auxiliary covariates. The auxiliary covariates include demographic, diagnosis, medication, and clinical-event features. Continuous variables are normalized using training-set parameters, and binary variables are appropriately encoded before model fitting.

We use fixed patient-level partitions of 56% training, 14% validation, and 30% test data across all methods. The validation set is used for model selection and the test set only for final evaluation. Cohort and partition sizes are summarized in Table 7.

Table 7: Cerner cohort and patient-level data partitions.

Partition	# Rows	$Y = 1$	$Y = 0$	Share
All	106,261	26,089	80,172	100%
Train	59,505	14,609	44,896	56%
Valid	14,877	3,653	11,224	14%
Test	31,879	7,827	24,052	30%

For semi-synthetic evaluation, we additionally mask originally observed laboratory values at rates of 5%, 10%, 30%, 50%, 70%, 90%, 95%, and 99%. Imputation accuracy is evaluated on these masked entries using their retained ground-truth values, with the patient partitions unchanged.

C.4 Implementations and Hyperparameters

We tune the five coefficients in the expanded objective Eq. 10. The search spaces and selected values are reported in Table 8.

PU learning. For each target variable j, C_{ψ_j} and D_{ω_j} are two-layer MLPs with hidden widths 64 and 32, batch normalization, LeakyReLU, and dropout. Both use the 157-dimensional $\mathbf{a}_i$ as input, and separate classifier–discriminator pairs are trained for the 167 targets. Optimization uses Adam with learning rate 10^{-3}, weight decay 10^{-4}, and minibatches of 128 positive and 128 unlabeled instances. The PAN adversarial coefficient is selected from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ using the validation set.

VAE-MNAR. The model uses 24-dimensional feature embeddings, a 16-dimensional Gaussian latent space, and two-layer encoder and decoder networks with widths 128 and 64. Continuous and

Table 8: Search spaces and selected loss coefficients.

Coefficient	Search range	Scale	Selected
λ_{cont}	$\{1, 2, 3\}$	Discrete	2
λ_{bin}	$\{0.25, 0.5, 1\}$	Discrete	0.5
λ_M	$[1, 3]$	Linear	2.80
λ_{KL}	$[0.01, 0.1]$	Log	0.022
λ_{PU}	$[10^{-4}, 0.1]$	Log	0.036

binary variables are modeled with Gaussian and Bernoulli likelihoods, respectively. The missingness model uses hidden widths 64 and 32, and the conditional Gaussian prior is parameterized by $(\mathbf{a}_i, \mathbf{q}_i^*)$. Training uses 20 importance samples per observation and Adam (Kingma & Ba, 2015) with learning rate 10^{-3}.

C.5 Implementations of baselines

We implement most baselines using author-provided implementations and publicly available code.

- **MICE** (Buuren & Groothuis-Oudshoorn, 2011) — `https://www.jstatsoft.org/article/view/v045i03`
- **MissForest** (Stekhoven & Bühlmann, 2012) — `https://github.com/stekhoven/missForest`
- **MisGAN** (Li et al., 2019) — `https://github.com/steveli/misgan`
- **GAIN** (Yoon et al., 2018) — `https://github.com/jsyoon0823/GAIN`
- **GINA** (Ma & Zhang, 2021) — `https://github.com/microsoft/project-azua`
- **not-MIWAE** (Ipsen et al., 2021) — `https://github.com/nbip/notMIWAE`
- **PSMVAE** (Ghalebikesabi et al., 2021) — `https://github.com/sghalebikesabi/PSMVAE`

For GNR (Chen et al., 2023), which does not provide public official codebase, we use our own implementation based on the original paper.

C.6 Hyperparameter settings of baselines

We configure the following baselines with reference to the settings described in their original papers and accompanying documentation.

- **MICE** (Buuren & Groothuis-Oudshoorn, 2011): we use the default settings documented for the `mice()` function.[1]
- **MissForest** (Stekhoven & Bühlmann, 2012): we follow the random-forest parameter settings reported in Section 4.5 (Computational efficiency) of the original paper.
- **MisGAN** (Li et al., 2019): we refer to Appendix D (Architectural details and hyperparameters) of the original paper for the training hyperparameters.
- **GAIN** (Yoon et al., 2018): we follow the hyperparameter specification in Section 3 (Hyperparameters) of the supplementary material.
- **GINA** (Ma & Zhang, 2021): we refer to Appendix B (Implementation details) of the original paper for the model and training settings.
- **not-MIWAE** (Ipsen et al., 2021): we follow the tabular-data settings described in Appendix A.1 (UCI), together with the general optimization settings in Appendix A.
- **PSMVAE** (Ghalebikesabi et al., 2021): we follow the model and training settings in Section 2.2 (Implementation Details) of the supplementary material.

[1] `https://amices.org/mice/reference/mice.html`