# A Double Projection Approach for Safe and Efficient Semi-Supervised Data-Fusion

Yiming Li*, Xuehan Yang*, Ying Wei, Molei Liu

Feb 2023

**Abstract**

Advances in data collection and transmission technologies have made larger amounts of data readily available. However, there are differences in the data collection capabilities of different data centers, or there are inevitable data missing. Many previous approaches to handling missing information have solely focused on either missing predictors or missing responses. In this paper, we will consider both types of missing and incorporate more information by projecting score functions into subsets, thus proposing algorithms that have ensured efficiency relative to the complete-case analysis. By generalizing the algorithm of this paper, it is promising to be able to handle more complex missing data structures in the future.

## 1 Introduction

We encounter multisource or multimodality data frequently in many real data applications. One common example might be electronic health record (EHR) systems adopted by most health care and medical facilities nowadays. However, blocks of variable information could be completely missing as there might be no need or it might be infeasible to collect certain sources of information given other known variables. The most common method for handling missing data is to perform complete-case analysis which removes observations with missing values. However, complete-case analysis heavily narrows the available and suitable population. Thus, Inverse Probability Weight (IPW), Imputation, Generative Adversarial Network (GAN), and Diffusion have been proposed. But most of the above methods focus on the case that either predictors are partially missing or unlabeled data while not taking structure block-wise missing into consideration. We proposed an integrated statistical method with guaranteed efficiency that combines both predictors missing and response missing simultaneously. Our proposed method allows more unlabeled data and partially incomplete data to be included and naturally enhances the utilization of potential data. We also show that many consistent machine learning methods could be applied in the estimation while the target coefficients are still bounded by $o_p(n^{-\frac{1}{2}})$. **not enough**

*The two authors have equal contributions to this work.

## 1.1 Literature Review

## 1.2 Main Contribution

In this study, under carefully constructed unbiased estimation equations and considering the more complex missing structure, we successfully proposed a two-stage estimator with guaranteed efficiency for low-dimension logistic regression coefficient vectors and obtained its theoretical properties under mild regularity conditions for smoothness and derivability. Significantly, unlike the existing method, our proposed method considers labeled-incomplete and unlabeled-complete data sequentially and we adroitly introduce the projection to deal with the missingness. In addition, we also prove that our proposed estimators have nice asymptotic properties and only require the consistency of nuisance parameters for those imputation methods applied in the intermediate state. Since our proposed estimators are sequential, our theoretical justifications further prove that the utilization of unlabeled-complete data and labeled-incomplete data have sequential efficiency and closed forms for variance reductions.

# 2 Preliminaries

## 2.1 Data Structure

Our interest lies in evaluating a prediction model for a binary response $Y$ based on a predictor vector $\boldsymbol{X} = (x_1, x_2, ..., x_p)^\top$ for some fixed dimension p. The underlying full data consists of $N = N_{\mathcal{LC}} + N_{\mathcal{UC}} + N_{\mathcal{LM}}$ independent and identically distributed random vectors.

$$\mathcal{D} = \{\boldsymbol{D}_i = (Y_i, \boldsymbol{X}_i^\top, t_i)^\top\}_{i=1}^N$$

where $t_i \in T = \{\mathcal{LC}, \mathcal{UC}, \mathcal{LM}\}$ is a discrete stratification variable that defines the data type: $\mathcal{LC}$ represents the **labeled complete** data. $\mathcal{UC}$ represents the **unlabeled complete** data. $\mathcal{LM}$ represents the **labeled incomplete** data ($\mathcal{M}$ represents missing) with some missing covariates and $\mathcal{X}_{\mathcal{LM}} = \{\boldsymbol{X}_i : t_i = \mathcal{LM}\}$.

We start with the easier case by assuming all $\boldsymbol{X}_i \in \mathcal{X}_{\mathcal{LM}}$ miss the same positions, denoted by $\mathcal{P}_m = \{j : x_j \text{ is missed}\}$ and $\mathcal{P}_{m^c} = \{1, 2, ..., p\}/\mathcal{P}_m$. Without loss of generality, we can suppose $\mathcal{P}_m = \{p - |\mathcal{P}_m| + 1, p - |\mathcal{P}_m| + 2, ..., p\}$ which means that all the missing data is placed in the latter part. We further suppose that the first $N_{\mathcal{LC}}$ data is labeled complete, the next $N_{\mathcal{LM}}$ data is labeled incomplete, and the last $N_{\mathcal{UC}}$ data is unlabeled complete.

Figure 1 is a brief visualization of the data structure. In the real health data application, $\boldsymbol{X}$ usually represents the measurements, lab test results, demographics, and so on, which could be easily obtained if the equipment meets the requirements. $Y$ usually represents the diagnosis of diseases, which requires a comprehensive diagnosis by multiple professionals. $Y$ could also be some golden labels, which require strict criteria to make the conclusion. Thus, the collection of $\mathcal{LC}$ and $\mathcal{LM}$ would be more difficult than the collection of $\mathcal{UC}$. We then assume that $N_{\mathcal{UC}}$ is quite larger than $N_{\mathcal{LC}}, N_{\mathcal{LM}}$.

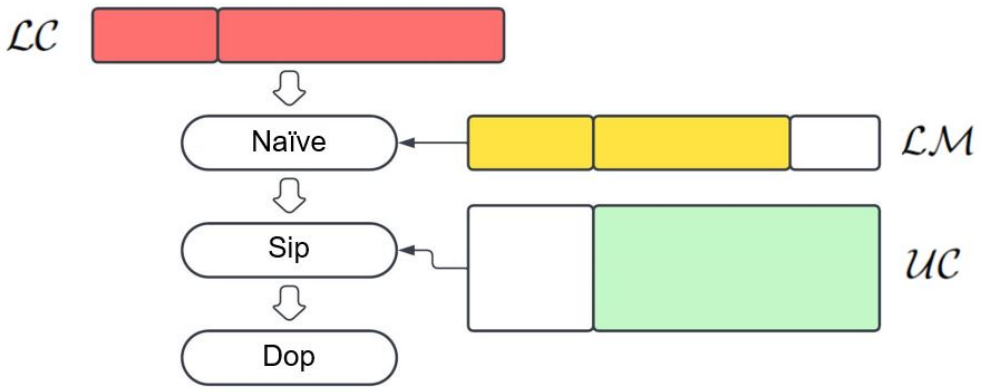


Figure 1: Data Structure

## 2.2 Problem Setup

To predict $Y$ based on $\boldsymbol{X}$, we fit a logistic regression model

$$\mathrm{P}(Y = 1|\boldsymbol{X}) = g(\boldsymbol{\gamma}^\top \boldsymbol{X})$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_p)^\top$ is an unknown vector of regression parameters and $g(\cdot)$ mapping $(-\infty, \infty) \to (0, 1)$ is a specified, smooth monotone function such as the expit function. The target (true) model parameter, $\bar{\boldsymbol{\gamma}}$, is the solution to the estimating equation:

$$\mathbf{U}(\boldsymbol{\gamma}) = \mathbb{E}[\boldsymbol{X}\{Y - g(\boldsymbol{\gamma}^\top \boldsymbol{X})\}] = \mathbf{0}$$

The easiest way of obtaining an estimation of the true parameter is to adopt the complete case analysis and get what we call LC-only estimator $\widetilde{\boldsymbol{\gamma}}$, which is the solution to the estimating equation:

$$\widetilde{\mathbf{U}}_{N_{\mathcal{LC}}}(\boldsymbol{\gamma}) = \frac{1}{N_{\mathcal{LC}}} \sum_{t_i = \mathcal{LC}} \boldsymbol{X}_i\{Y_i - g(\boldsymbol{\gamma}^\top \boldsymbol{X}_i)\} = \mathbf{0}$$

$\widetilde{\boldsymbol{\gamma}}$ is the standard Z-estimation. Thus, under some standard and mild constraints, we declare that $\widetilde{\boldsymbol{\gamma}}$is a regular root-$N_{\mathcal{LC}}$ consistent estimator for the unique solution.

In this paper, we aim to obtain a sequence of semi-supervised estimators: the advanced estimator utilizing projections on labeled incomplete data is called single projection **SiP** estimator with notation $\widetilde{\boldsymbol{\gamma}}_1$ and the advanced estimator utilizing projections on labeled incomplete data then projections on unlabeled complete data is called double projection **DoP** estimator with notation $\widetilde{\boldsymbol{\gamma}}_2$. Both of them should be asymptotic variance efficient compared to the LC-only estimator $\widetilde{\boldsymbol{\gamma}}$.

We focus on two additional data sources with respect to $\mathcal{LC}$ data: $\mathcal{X}_{\mathcal{LM}}$ and $\mathcal{X}_{\mathcal{UC}}$ sequentially as they convey distinct information about the likelihood of the prediction model under the i.i.d. assumptions we have made. We also hope Sip and Dop are root-$N_{\mathcal{LC}}$ consistent estimators. However, the construction of Sip and Dop considers the data fusion idea, thus them does not follow the rule of standard M-estimation or Z-estimation. We show in the Asymptotic section that our proposed Sip and Dop estimators are root-$N_{\mathcal{LC}}$ consistent.

# 3 Estimation Procedure

## 3.1 Step 1: LC-only Estimation

We hope to get a basic estimation of the true parameter. To catch this, the LC-only estimator is calculated based on labeled complete data $\mathcal{LC}$. Since our aim is to compare the variance of the LC-only estimator and our proposed Sip and Dop estimators, we hope to record the variance of the LC-only estimator. Further, the asymptotic variance of an estimator is determined by the Fisher information, which is also the variance of the score function. We thus record the score function of the LC-only estimator to better capture the variance.

We first construct a generalized linear model to catch the relationship between $Y$ and $\boldsymbol{X}$ by solving the following equation:

$$\widetilde{\mathbf{U}}_{N_{\mathcal{LC}}}(\boldsymbol{\gamma}) = \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \boldsymbol{X}_i\{Y_i - g(\boldsymbol{\gamma}^\top \boldsymbol{X}_i)\} = \mathbf{0} \tag{1}$$

The purpose of the regression is to make use of the labeled complete data and give an LC-only estimation $\widetilde{\boldsymbol{\gamma}}$ of the true target model parameter $\bar{\boldsymbol{\gamma}}$. We note that $\widetilde{\boldsymbol{\gamma}}$ is a standard Z-estimator, by **A.W. VANDER VAART**, the LC-only estimator on labeled complete data has the unique solution and satisfies the following equation:

$$\widetilde{\boldsymbol{\gamma}} = \bar{\boldsymbol{\gamma}} + \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \boldsymbol{H}^{-1}\boldsymbol{X}_i\{Y_i - g(\bar{\boldsymbol{\gamma}}^\top \boldsymbol{X}_i)\} + o_p(\frac{1}{\sqrt{N_{\mathcal{LC}}}})$$

where hessian matrix $\bar{\boldsymbol{H}}$ is defined as $\bar{\boldsymbol{H}} = \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \dot{g}(\bar{\boldsymbol{\gamma}}^\top \boldsymbol{X}_i)\boldsymbol{X}_i\boldsymbol{X}_i^\top$ and score function $\boldsymbol{S}_i$ is defined as $\bar{\boldsymbol{S}}_i = \bar{\boldsymbol{H}}^{-1}\boldsymbol{X}_i\{Y_i - g(\bar{\boldsymbol{\gamma}}^\top \boldsymbol{X}_i)\}$. In the following step, we use the $\widetilde{\boldsymbol{H}}$ and $\widetilde{\boldsymbol{S}}$ to represent the numerical estimations of $\boldsymbol{H}$ and $\boldsymbol{S}$ by replacing $\bar{\boldsymbol{\gamma}}$ with $\widetilde{\boldsymbol{\gamma}}$, i.e., $\widetilde{\boldsymbol{H}} = \sum_{t_i=\mathcal{LC}} \dot{g}(\widetilde{\boldsymbol{\gamma}}^\top \boldsymbol{X}_i)\boldsymbol{X}_i\boldsymbol{X}_i^\top$ and $\widetilde{\boldsymbol{S}}_i = \widetilde{\boldsymbol{H}}^{-1}\boldsymbol{X}_i\{Y_i - g(\widetilde{\boldsymbol{\gamma}}^\top \boldsymbol{X}_i)\}$. The consistency of $\widetilde{\boldsymbol{H}}$ and $\widetilde{\boldsymbol{S}}$ is justified in the asymptotic property section. Under the condition 1-6 in the asymptotic section, we argue that LC-only estimator $\widetilde{\boldsymbol{\gamma}}$ is root-$N_{\mathcal{LC}}$ consistent. When the LC-only model is correctly specified, $\widetilde{\boldsymbol{\gamma}}$ gives a pretty accurate approximation of the true parameter. However, we tend to get an unstable estimation which highly depends on the sample size and potential distribution of data when the model is misspecified. We increase the sample by including imputed value on data $\mathcal{LM}$ and $\mathcal{UC}$ and cross-fitting to solve this question.

**On the consequences of model misspecification in logistic regression**

## 3.2 Step 2: SiP Estimaton

We hope to aggregate $\mathcal{LM}$ data to improve the performance of our Sip estimator. The main problem we meet is how to make up the missingness of $\boldsymbol{X}_{\mathcal{P}_m}$ since the score functions, determinants of the asymptotic variance, rely on the observed value of each column. It is natural to consider the projection of the score function on $\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y$ (defined as $\widetilde{\varphi}_1$ in the equation 2). Thus, the conditional distribution $\boldsymbol{X}_{\mathcal{P}_m} \mid (\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y)$ is needed, and this conditional distribution should be the same across the whole population. Then we estimate the density of $\boldsymbol{X}_{\mathcal{P}_m} \mid (\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y)$ utilizing the labeled complete data $\mathcal{LC}$. Since $\widetilde{\varphi}_1$'s no more

depends on $\boldsymbol{X}_{\mathcal{P}_m}$, it's accessible to $\mathcal{LM}$. Finally, we could construct our Sip estimator by aggregating more $\widetilde{\varphi}_1$'s in $\mathcal{LM}$ data.

We start by considering the conditional density $P(\boldsymbol{X}_{\mathcal{P}_m} \mid \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y)$ on $\mathcal{LC}$ data. Recall the definition of the score function:

$$\bar{\boldsymbol{S}}_i = \bar{\boldsymbol{H}}^{-1}\boldsymbol{X}_i\{Y_i - g(\bar{\boldsymbol{\gamma}}^\top \boldsymbol{X}_i)\},$$

instead of simply imputing the $\boldsymbol{S}$ on $\mathcal{LM}$, we project the score function on the subset $\{\boldsymbol{X}_{\mathcal{P}_m}, Y\}$ of $\{\boldsymbol{X}, Y\}$. We incorporate conditional density to give an unbiased estimator of the score function: $\boldsymbol{\varphi}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y) = \mathbb{E}[\boldsymbol{S} \mid \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y]$. This is the step we called Single Projection. We construct its numerical estimation through the following equations:

$$\widetilde{\boldsymbol{\varphi}}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y = 0, 1) = \int \widetilde{\boldsymbol{S}}\widetilde{P}(\boldsymbol{X}_{\mathcal{P}_m} \mid \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y = 0, 1)\, d\boldsymbol{X}_{\mathcal{P}_m} \tag{2}$$

Note that $\widetilde{P}(\boldsymbol{X}_{\mathcal{P}_m} \mid \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y = 0, 1)$ can be estimated through many potential procedures, such as (multivariate) linear regression, non-parametric approximation. We show in the Asymptotic Property section that if distributions calculated from these procedures are consistent with the true distribution, we could have a root-n Sip estimator. Generally, this requirement is easy to attach.

Another thing we need to mention is the calculation of integration. Considering the following two points: some machine learning models have no explicit form of distribution; the integration might not have a closed form. We use the Monte Carlo method to numerically approximate the true integration with the belief that the Monte Carlo integration would converge to the true one when the number of replicates increases, which is equivalent to saying the Monte Carlo integration is consistent.

We first consider the (multivariate) linear regression of $\boldsymbol{X}$. The mean of $\boldsymbol{X}_{\mathcal{P}_m}$ equals the linear combination of the coefficients and $\boldsymbol{X}_{\mathcal{P}_{m^c}}$ and $Y$. We also give an estimation of variance by assuming that the variance is homogeneous among all the data and solve it by maximizing the likelihood. Thus

$$\boldsymbol{X}_{i\mathcal{P}_m} = \beta_1^\top \widetilde{\boldsymbol{X}}_{i\mathcal{P}_{m^c}} + \beta_2 Y + \boldsymbol{\epsilon} \quad \mathrm{Var}(\boldsymbol{\epsilon}) = \frac{1}{N_{\mathcal{LC}}} \sum_{t_i = \mathcal{LC}} [(\boldsymbol{X}_{i\mathcal{P}_m} - \widetilde{\boldsymbol{X}}_{i\mathcal{P}_m})^\top (\boldsymbol{X}_{i\mathcal{P}_m} - \widetilde{\boldsymbol{X}}_{i\mathcal{P}_m})]$$

where $\widetilde{\boldsymbol{X}}_{i\mathcal{P}_m}$ is the predicted value of $\boldsymbol{X}_{i\mathcal{P}_m}$ through the (multivariate) linear model we have constructed and $\boldsymbol{X}_{i\mathcal{P}_m} \sim \mathcal{N}(\widetilde{\boldsymbol{X}}_{i\mathcal{P}_m}, \mathrm{Var}(\boldsymbol{\epsilon}))$. Now we calculate $\widetilde{\boldsymbol{\varphi}}_1$ defined in the equation 2.

Sometimes it's not so accurate to simply replace the score function with $\boldsymbol{\varphi}_1$ since $\boldsymbol{\varphi}_1$ is only the partial projection of the score function. To solve this, we hope to further approach each dimension of the score function utilizing $\boldsymbol{\varphi}_1$. We begin with linear regression between $\boldsymbol{S}_i(j)$ and $\boldsymbol{\varphi}_1$ where $\boldsymbol{S}_i(j)$ is the jth element of $\boldsymbol{S}_i$. Then we consider the linear regression between $\boldsymbol{h}^\top \boldsymbol{S}$ and $\boldsymbol{\varphi}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y)$ where $\boldsymbol{h}$ is an arbitrary vector with the same dimension as $\boldsymbol{S}$. If we could find the way how $\boldsymbol{S}$ is influenced by $\widetilde{\boldsymbol{\varphi}}_1$, then we are able to reconstruct the LC-only estimator to the augmented estimator. The reason we consider the linear regression method is that the residual and the predictors are orthogonal with each other. This nice property makes it convenient to organize the variance of our proposed efficient **SiP** estimator. Suppose the target model parameter, $\bar{\boldsymbol{\alpha}}_1$, is the solution to the equation:

$$\mathbf{U}_{N_{\mathcal{LC}}}(\boldsymbol{\alpha}_1) = \mathbb{E}\boldsymbol{\varphi}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y)\{\boldsymbol{h}^\top \boldsymbol{S} - \boldsymbol{\alpha}_1^\top \boldsymbol{\varphi}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y)\} = \mathbf{0}$$

Then the $\widetilde{\boldsymbol{\alpha}}_1$ is the solution to the following numerical equation:

$$\widetilde{\mathbf{U}}_{N_{\mathcal{LC}}}(\boldsymbol{\alpha}_1) = \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\varphi}}_1(\boldsymbol{X}_{i\mathcal{P}_{m^c}}, Y_i)\{\boldsymbol{h}^\top \widetilde{\boldsymbol{S}}_i - \boldsymbol{\alpha}_1^\top \widetilde{\boldsymbol{\varphi}}_1(\boldsymbol{X}_{i\mathcal{P}_{m^c}}, Y_i)\} = \mathbf{0}$$

We also assume that all individuals are homogeneous, so the expectation of $\boldsymbol{\varphi}_1$ should be the same on either data $\mathcal{LC}$ or data $\mathcal{LM}$. Thus, we proposed the augmented model parameter $\widetilde{\boldsymbol{\gamma}}_1$ through the following equation:

$$\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_1 = \boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}} - \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i} + \frac{1}{N_{\mathcal{LC}} + N_{\mathcal{LM}}} \sum_{t_i=\mathcal{LC},\mathcal{LM}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i} \tag{3}$$

We show in the latter section that the proposed SiP estimator $\widetilde{\boldsymbol{\gamma}}_1$ is fully efficient compared to the naive estimator $\widetilde{\boldsymbol{\gamma}}$ and unbiased to true model parameter $\bar{\boldsymbol{\gamma}}$. The other thing we want to emphasize is why we first consider the $\mathcal{LM}$ data instead of $\mathcal{UC}$ data. Remember we obtain $\widetilde{\boldsymbol{\varphi}}_1$ by projecting the score function a subset $\boldsymbol{X}_{\mathcal{P}_{m^c}}$ and $Y$. However, this is unreliable when we consider the $\mathcal{UC}$ data since $\boldsymbol{S}$ is orthogonal with $\boldsymbol{X}$. Another important thing is that Sip estimator should be unbiased to the LC-only estimator since $\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}$ and $\frac{1}{N_{\mathcal{LC}}+N_{\mathcal{LM}}} \sum_{t_i=\mathcal{LC},\mathcal{LM}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}$ are both numerical estimation of $\mathbb{E}\widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}$. Thus, the difference between these two should converge to zero when sample sizes $N_{\mathcal{LC}}$ and $N_{\mathcal{LM}}$ increase. Intuitively, we might believe Sip estimator should outperform LC-only estimator since we include more information in $\mathcal{LM}$ data.

### 3.3 Step 3: Dop Estimation

We hope to aggregate $\mathcal{LM}$ data to improve the performance of our Sip estimator so we further consider the projection of score functions on $\boldsymbol{X}$. However, as we have mentioned before, $\boldsymbol{X}$ and score functions are inherently orthogonal with each other. Thus, we consider the projection of $\widetilde{\varphi}_1$'s on $\boldsymbol{X}$ (defined as $\widetilde{\varphi}_2$ in the equation 4). We construct our Dop estimator by aggregating more $\widetilde{\varphi}_2$'s in $\mathcal{UC}$ data.

In the above procedure, we consider the projection of $\boldsymbol{S}$ and obtain $\boldsymbol{\varphi}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y)$. The proposed single projection $\boldsymbol{\varphi}_1$ is no longer orthogonal with $\boldsymbol{X}$ and we could apply the second projection and get the **DoP** of the score function. In this step, we consider the projection of $\boldsymbol{\varphi}_1$ conditioning on $\boldsymbol{X}$, i.e., $\mathbb{E}[\boldsymbol{\varphi}_1 \mid \boldsymbol{X}]$. $\boldsymbol{\varphi}_1$ is the function of $\boldsymbol{X}_{\mathcal{P}_{m^c}}$ and $Y$. $\boldsymbol{X}_{\mathcal{P}_{m^c}}$ is part of $\boldsymbol{X}$ so $\boldsymbol{X}_{\mathcal{P}_{m^c}} \mid \boldsymbol{X} = \boldsymbol{X}_{\mathcal{P}_{m^c}}$. Thus we only need to consider the distribution of $Y \mid \boldsymbol{X}$. This conditional density of $Y$ could be derived from equation (1), denoted by $P(Y \mid \boldsymbol{X})$. Then we construct the projection of $\boldsymbol{\varphi}_1$ through the follow equation:

$$\begin{aligned}\boldsymbol{\varphi}_2(\boldsymbol{X}) &= \int \boldsymbol{\varphi}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y) P(Y \mid \boldsymbol{X})\, dY \\ &= P(Y=1 \mid \boldsymbol{X})\boldsymbol{\varphi}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y=1) + P(Y=0 \mid \boldsymbol{X})\boldsymbol{\varphi}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y=0)\end{aligned}$$

We also have numerical estimations of $\boldsymbol{\varphi}_2$:

$$\widetilde{\boldsymbol{\varphi}}_2(\boldsymbol{X}) = \widetilde{P}(Y=1|\boldsymbol{X})\widetilde{\boldsymbol{\varphi}}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y=1) + \widetilde{P}(Y=0|\boldsymbol{X})\widetilde{\boldsymbol{\varphi}}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y=0) \tag{4}$$

Notice that $Y$ is the binary random variable, so its distribution can be separated into two parts. In the Sip estimation, we only record $\boldsymbol{\varphi}_1$ for the exact observed $Y$ (either 0 or 1). Here we need to calculate the virtual $\boldsymbol{\varphi}_1$ for $Y$ (both 0 and 1). Then we could calculate $\boldsymbol{\varphi}_2$ based on the probability obtained in the LC-only estimation.

Recall we include the information of the $\mathcal{LM}$ data when constructing the current model parameter $\widetilde{\boldsymbol{\gamma}}_1$ by utilizing the single projection of score functions. We also hope to use double projections and then include the information on the $\mathcal{UC}$ data. To keep the orthogonality between the $\mathcal{LC}$ part and the $\mathcal{UC}$ part, we consider the following linear regression and suppose the target model parameter, $\bar{\boldsymbol{\alpha}}_2$, is the solution to the equation:

$$\mathbf{U}_{N_{\mathcal{LC}}}(\boldsymbol{\alpha}_2) = \mathbb{E}\boldsymbol{\varphi}_2(\boldsymbol{X})\{\boldsymbol{h}^\top \boldsymbol{S} - \frac{N_{\mathcal{LM}}}{N_{\mathcal{LM}} + N_{\mathcal{LC}}}\boldsymbol{\alpha}_1^\top \boldsymbol{\varphi}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y) - \boldsymbol{\alpha}_2^\top \boldsymbol{\varphi}_2(\boldsymbol{X})\} = \mathbf{0}$$

And its numerical approximation is the solution to the following estimation equation:

$$\widetilde{\mathbf{U}}_{N_{\mathcal{LC}}}(\boldsymbol{\alpha}_2) = \frac{1}{N_{\mathcal{LC}}}\sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\varphi}}_2(\boldsymbol{X}_i)\{\boldsymbol{h}^\top \widetilde{\boldsymbol{S}}_i - \frac{N_{\mathcal{LM}}}{N_{\mathcal{LM}} + N_{\mathcal{LC}}}\widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_1(\boldsymbol{X}_{i\mathcal{M}^c}, Y_i) - \boldsymbol{\alpha}_2^\top \widetilde{\boldsymbol{\varphi}}_2(\boldsymbol{X}_i)\} = \mathbf{0}$$

Where $\boldsymbol{h}^\top \boldsymbol{S} - \frac{N_{\mathcal{LM}}}{N_{\mathcal{LM}}+N_{\mathcal{LC}}}\boldsymbol{\alpha}_1^\top \boldsymbol{\varphi}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y)$ is the term related to $\mathcal{LC}$ data after the construction of the Sip estimator. Since we also assume that all individuals are homogeneous, the expectation of $\boldsymbol{\varphi}_2$ should be the same on either data $\mathcal{LC}$ or data $\mathcal{UC}$. Thus, we proposed the augmented model parameter $\widetilde{\boldsymbol{\gamma}}_2$ through the following equation:

$$\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_2 = \boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_1 - \frac{1}{N_{\mathcal{LC}}}\sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i} + \frac{1}{N_{\mathcal{LC}} + N_{\mathcal{UC}}}\sum_{t_i=\mathcal{LC},\mathcal{UC}} \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i} \tag{5}$$

We show in the latter section that the proposed estimator $\widetilde{\boldsymbol{\gamma}}_2$ is fully efficient and unbiased to $\widetilde{\boldsymbol{\gamma}}_1$ due to the nice property of linear regression.

## 3.4 Pesudo Code of Algorithm and Cross fitting method

- Step 1: Solve $\widetilde{\boldsymbol{\gamma}}$ from logistic regression based on $Y, \boldsymbol{X}$ in the $\mathcal{LC}$ data.
- Step 2: Randomly divide the $\mathcal{LC}$ data into K even folds, and calculate $\mathbb{E}^{(k)}[\boldsymbol{S}(\boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y]$ based on data $\mathcal{LC}^{(-k)}$. The calculate $\boldsymbol{\varphi}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y) = \mathbb{E}[\boldsymbol{S}(\boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y] = \frac{1}{K}\sum_{k=1}^K \mathbb{E}^{(k)}[\boldsymbol{S}(\boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y]$ in both $\mathcal{LC}$ data and $\mathcal{LM}$ data. Calculate $\widetilde{\boldsymbol{\alpha}}_1$ by solving the linear regression $\boldsymbol{h}^\top \widetilde{\boldsymbol{S}} \sim \widetilde{\boldsymbol{\varphi}}_1$, $\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_1 = \boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}} - \frac{1}{N_{\mathcal{LC}}}\sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i} + \frac{1}{N_{\mathcal{LC}}+N_{\mathcal{LM}}}\sum_{t_i=\mathcal{LC},\mathcal{LM}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}$
- Step 3: Calculate $\boldsymbol{\varphi}_2(\boldsymbol{X}) = \mathbb{E}[\boldsymbol{\varphi}_1 \mid \boldsymbol{X}]$ in both $\mathcal{LC}$ data and $\mathcal{UC}$ data. Calculate $\widetilde{\boldsymbol{\alpha}}_2$ by solving the linear regression $\boldsymbol{h}^\top \widetilde{\boldsymbol{S}} - \frac{N_{\mathcal{LM}}}{N_{\mathcal{LM}}+N_{\mathcal{LC}}}\widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_1 \sim \widetilde{\boldsymbol{\varphi}}_2$, $\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_2 = \boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_1 - \frac{1}{N_{\mathcal{LC}}}\sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i} + \frac{1}{N_{\mathcal{LC}}+N_{\mathcal{UC}}}\sum_{t_i=\mathcal{LC},\mathcal{UC}} \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i}$

Currently, we have obtained our proposed sequential two-stage projection estimators. We learn conditional distributions of some specific variable to give estimations of projections and distribution learning could be highly dependent on the potential population. When $\boldsymbol{X}_{\mathcal{P}_m} \mid \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y$ is correctly specified, Sip estimation returns a more precise estimation of

$\widetilde{\boldsymbol{\varphi}}_1$. However, misspecified models are not uncommon in real cases. Our proposed estimators use the data $\mathcal{LC}$ to construct and evaluate the prediction model and are therefore prone to overfitting bias especially when utilizing the machine learning method [**Efron, 1986**]. The reason is that machine learning methods use regularization to decrease the variance of an estimator. However, there is a trade-off between an introduced bias on the parameter of interest through regularization and over-fitting. Cross-fitting is a practical, efficient form of data splitting to solve this issue. Importantly, its use here is not simply as a device to make proofs elementary, but as a practical method to allow us to overcome the overfitting/high-complexity phenomena that commonly arise in data analysis based on highly adaptive ML methods.[**Double/debiased machine learning for treatment and structural parameters**]. Here, we propose an averaged K-fold cross-fitting estimator.

Suppose the $\mathcal{LC}$ data is randomly split into $K$ different samples with roughly equal sample size, denoted by $\{\mathcal{LC}^{(1)}, \mathcal{LC}^{(2)}, ..., \mathcal{LC}^{(K)}\}$, where $|\mathcal{LC}^{(k)}| \approx \frac{n}{k}$ and $\mathcal{LC}^{(-k)} = \{\mathcal{LC}^{(j)} : j \neq k\}$ for $k \in [1 : K]$. For each $k$, apply machine learning methods such as random forest or kernel smooth to regress $\boldsymbol{X}_{\mathcal{P}_m} \sim \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y$ with respective to nuisance estimator $\bar{\boldsymbol{\eta}}$, denoted as $\widetilde{\boldsymbol{\eta}}^{(-k)}$ based on the data $C^{(-k)}$. For the k-th fold, we estimate $\boldsymbol{X}_{i\mathcal{P}_m} \mid \boldsymbol{X}_{i\mathcal{M}^c}, Y_i$ based on $\widetilde{\boldsymbol{\eta}}^{(-k)}$ for each $i \in \mathcal{LM}$ and the results are recorded as $\widetilde{\boldsymbol{X}}^{(-k)}_{i\mathcal{M}} \mid \boldsymbol{X}_{i\mathcal{M}^c}, Y_i$. The average K-fold cross-fitting estimation $\widetilde{\boldsymbol{X}}^{CF}_{i\mathcal{P}_m} \mid \boldsymbol{X}_{i\mathcal{P}_{m^c}}, Y_i = \frac{1}{K}\sum_{k=1}^{K} \widetilde{\boldsymbol{X}}^{(-k)}_{i\mathcal{P}_m} \mid \boldsymbol{X}_{i\mathcal{P}_{m^c}}, Y_i$ is adopted. We may similarly obtain a CF-based estimator, denoted by $\widetilde{\boldsymbol{\varphi}}_1^{CF}$.

We mainly consider two properties of the CF-based estimator. First, with the increases of fold counts k (k<n), the estimation seems to be less biased. This is the trade-off between bias and variance with respect to model complexity. Second, the CF-based estimator is still consistent with the true estimator, i.e., $\widetilde{\boldsymbol{\varphi}}_1^{CF} - \bar{\boldsymbol{\varphi}}_1$ and $\widetilde{\boldsymbol{\varphi}}_1 - \bar{\boldsymbol{\varphi}}_1$ are first-order asymptotically equivalent and bounded by $o_P(1)$. The proof is shown in the appendix. We still use $\widetilde{\boldsymbol{\varphi}}_1$ to simplify the illustration.

# 4 Asymptotic Property

In this section, we provide the theoretical justification of proposed sequential estimators $\widetilde{\boldsymbol{\gamma}}_1$, $\widetilde{\boldsymbol{\gamma}}_2$, and some preliminaries for nuisance parameters $\widetilde{\boldsymbol{\alpha}}_{1,2}$, $\widetilde{\boldsymbol{\varphi}}_{1,2}$. To facilitate our presentation, we first discuss the properties of $\widetilde{\boldsymbol{\gamma}}$ as the initial parameter to quantify the variability in estimating $\bar{\boldsymbol{\gamma}}$. We then present our main result highlighting the efficiency gain of our proposed sequential approach for accurate parameter estimation. We conclude our theoretical analysis with practical discussions of the efficiency of our proposed estimator in the previous setting.

Also, we make more notations to simplify the illustration. For our asymptotic analysis, $P(\boldsymbol{X}, \boldsymbol{\gamma})$ represents the probability density function of $\boldsymbol{X}$ with parameters $\boldsymbol{\gamma}$, $\boldsymbol{X}^{\otimes 2} = \boldsymbol{X}\boldsymbol{X}^\top$, $\mathbf{U}(\boldsymbol{\gamma}) = \mathbb{E}[\boldsymbol{X}\{Y - g(\boldsymbol{\gamma}^\top \boldsymbol{X})\}] = \mathbf{0}$. $\frac{N_{\mathcal{LM}}}{N_{\mathcal{LC}}} = \rho_1$, $\frac{N_{\mathcal{UC}}}{N_{\mathcal{LC}}} = \rho_2$. We also use $n$ to replace $N_{\mathcal{LC}}$. Before presenting the main theorems, we introduce the conditions for the proposed estimators.

**Condition 1.** $\mathbf{U}(\boldsymbol{\gamma})$ *has compact support and is of fixed dimension. The density function* $P(\boldsymbol{X}, \boldsymbol{\gamma})$ *of* $\boldsymbol{X}$ *is continuously differentiable. There is at least one continuous component of* $\boldsymbol{X}$ *corresponding non-zero component in coefficient* $\boldsymbol{\gamma}$

**Condition 2.** *(A)The link function* $g(\cdot)$ *is continuously differentiable with derivative* $\dot{g}(\cdot)$. *(B)*$\mathbf{0} < \mathbb{E}\left[\boldsymbol{X}^{\otimes 2}\dot{g}\left(\bar{\boldsymbol{\gamma}}^\top \boldsymbol{X}\right)\right] < \infty$.

**Condition 3.** *For any neighborhood of* $\bar{\boldsymbol{\gamma}}, \boldsymbol{\Theta} = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma} - \bar{\boldsymbol{\gamma}}\|_2 < \delta\}$ *with every* $\delta > 0$, $\inf_{\gamma \notin \boldsymbol{\Theta}} ||\mathbf{U}(\boldsymbol{\gamma})|| > 0$

**Remark.** *Conditions 1-3 are commonly used regularity conditions in zero-estimation theory and are satisfied in broad applications. Condition 3 assumes that there is no* $\boldsymbol{\gamma}$ *in the neighborhood of true* $\bar{\boldsymbol{\gamma}}$ *such that zero estimation is 0, which ensures the existence and uniqueness of the GLM estimator.* $\boldsymbol{\gamma}$*'s value at infinity also does not make that zero estimation equal to 0. Condition 1-2 guarantees the derivability of the aimed function and non-singularity of the designed Hessian matrix.*

The asymptotic properties of LC-only Estimator $\widetilde{\boldsymbol{\gamma}}$ are summarized in Theorem 1 and the justification is provided in the Appendix.

**Theorem 1.** *Under Condition 1–3, and* $\widetilde{\boldsymbol{\gamma}} \xrightarrow{p} \bar{\boldsymbol{\gamma}}$*, and*

$$\sqrt{n}(\widetilde{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}) = \boldsymbol{H}_{\bar{\boldsymbol{\gamma}}}^{-1} \frac{1}{\sqrt{n}} \sum_{t_i = \mathcal{LC}} \Phi_i + o_P(1)$$

*where*

$$\Phi_i = \boldsymbol{X}_i \{Y_i - g(\bar{\boldsymbol{\gamma}}^\top \boldsymbol{X}_i)\}$$
$$\boldsymbol{H}_{\bar{\boldsymbol{\gamma}}} = \mathbb{E}\left[\boldsymbol{X}^{\otimes 2} \dot{g}\left(\bar{\boldsymbol{\gamma}}^\top \boldsymbol{X}\right)\right]$$

*In particular, the sequence* $\sqrt{n}(\widetilde{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}})$ *is asymptotically normal with mean zero and covariance matrix* $\boldsymbol{H}_{\bar{\boldsymbol{\gamma}}}^{-1} \mathbb{E}[\Phi^{\otimes 2}] \boldsymbol{H}_{\bar{\boldsymbol{\gamma}}}^{-1\top}$

Condition 1-3 are commonly used for proof of the asymptotic properties. Recall we have defined $\bar{\boldsymbol{S}} = \bar{\boldsymbol{H}}^{-1} \boldsymbol{X}\{Y - g(\bar{\boldsymbol{\gamma}}^\top \boldsymbol{X})\}$, and we require the non-singularity of $\bar{\boldsymbol{H}}$, the covariance matrix of $\sqrt{n}(\widetilde{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}})$ can be expressed as $\mathbb{E}[\bar{\boldsymbol{S}}^{\otimes 2}]$. Since $\boldsymbol{H}$ and $\boldsymbol{S}$ are the functions of $\boldsymbol{\gamma}$ and the expectations of the first derivative are bounded, we claim that $\sqrt{n}(\widetilde{\boldsymbol{H}} - \bar{\boldsymbol{H}})$ and $\sqrt{n}(\widetilde{\boldsymbol{S}} - \bar{\boldsymbol{S}})$ are bounded by $C * o_P(1)$ where $C$ is expectation of the first derivative.

**Lemma 1.** $\widetilde{\boldsymbol{\varphi}}_1 \xrightarrow{p} \bar{\boldsymbol{\varphi}}_1$ $\widetilde{\boldsymbol{\alpha}}_1 \xrightarrow{p} \bar{\boldsymbol{\alpha}}_1$ *as* $n \to \infty$

For the analysis of $\widetilde{\boldsymbol{\varphi}}_1$, we consider the second-order Taylor expansion and get the interaction of $\widetilde{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\eta}}$. Even if these two parameters themselves might be asymptotic normal and independent, the product is not asymptotically normally distributed. In fact, the product of two independent normal distributions should follow the summation of two chi-square distributions. But we could still conclude that $\widetilde{\boldsymbol{\varphi}}_1$ is bounded by $o_p(1)$. The justification of $\widetilde{\boldsymbol{\varphi}}_1$'s consistency is detailed in the Appendix.
We also need to show $\widetilde{\boldsymbol{\alpha}}_1$ is bounded by $o_P(1)$. It is the coefficient from a linear regression, which should generally be $o_P(\frac{1}{\sqrt{n}})$, but error of $\widetilde{\boldsymbol{\alpha}}_1$ comes from two parts, one from the replacement of expectation with the sample mean, another from the error of $\widetilde{\boldsymbol{\varphi}}_1$. The proof for consistency of $\widetilde{\boldsymbol{\alpha}}_1$ should follow the proof of uniform consistency in **A.W. VANDER VAART** theorem 5.9. The justification in provided in the Appendix. Under lemma 1, we could conclude the asymptotic property of our Sip estimator.

**Theorem 2.**

$$\begin{aligned}&\sqrt{n}(\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_1-\boldsymbol{h}^\top\bar{\boldsymbol{\gamma}})\\&=\frac{\sum_{t_i=\mathcal{LC},\mathcal{LM}}\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_{1i}}{\sqrt{n}(1+\rho_1)}-\frac{\sum_{t_i=\mathcal{LC}}\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_{1i}}{\sqrt{n}}+\boldsymbol{h}^\top\boldsymbol{H}_{\bar{\boldsymbol{\gamma}}}^{-1}\frac{\sum_{t_i=\mathcal{LC}}\Phi_i}{\sqrt{n}}+o_P(1)\end{aligned}$$

*In particular, the sequence* $\sqrt{n}(\widetilde{\boldsymbol{\gamma}}_1-\bar{\boldsymbol{\gamma}})$ *is asymptotically normal with mean zero and covariance matrix* $\mathbb{E}[(\boldsymbol{h}^\top\boldsymbol{H}_{\bar{\boldsymbol{\gamma}}}^{-1}\Phi-\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^{\otimes 2}]+\frac{1}{1+\rho_1}\mathbb{E}[(\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^{\otimes 2}]$

We notice that there exists the product of two estimators and we want to show $\sqrt{n}(\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_1-\boldsymbol{h}^\top\bar{\boldsymbol{\gamma}})$ is still bounded by $o_P(1)$. The key step is to consider the expansion of $\widetilde{\boldsymbol{\alpha}}_1^\top\widetilde{\boldsymbol{\varphi}}_{1i}-\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_{1i}$, which includes two normal parts $\bar{\boldsymbol{\alpha}}_1^\top(\widetilde{\boldsymbol{\varphi}}_{1i}-\bar{\boldsymbol{\varphi}}_{1i})$, $\bar{\boldsymbol{\varphi}}_{1i}(\widetilde{\boldsymbol{\alpha}}_1^\top-\bar{\boldsymbol{\alpha}}_1^\top)$ and a non-normal part $(\widetilde{\boldsymbol{\alpha}}_1^\top-\bar{\boldsymbol{\alpha}}_1^\top)(\widetilde{\boldsymbol{\varphi}}_{1i}-\bar{\boldsymbol{\varphi}}_{1i})$. Thus, $\widetilde{\boldsymbol{\alpha}}_1^\top\widetilde{\boldsymbol{\varphi}}_{1i}-\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_{1i}$ is not asymptotic normal due to the interaction of term. However, we could conclude that $\sqrt{n}(\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_1-\boldsymbol{h}^\top\bar{\boldsymbol{\gamma}})$ is asymptotic normal with mean zero under the assumption that all the individuals have the same underlying distribution. This is because the bias is canceled since the expectations merely based on complete data or based on both complete and incomplete data should be the same, i.e., the first and second terms in theorem 2. Similarly, the error of $\widetilde{\boldsymbol{\gamma}}_1$ consists of two parts, errors from the naive supervised estimator and errors from modification term $(\widetilde{\boldsymbol{\alpha}}_1^\top\widetilde{\boldsymbol{\varphi}}_{1i})$. From theorem 1 we conclude that $\widetilde{\boldsymbol{\gamma}}$ is bounded by $o_P(n^{-\frac{1}{2}})$. Under our homogeneous assumption, $\widetilde{\boldsymbol{\alpha}}_1^\top\widetilde{\boldsymbol{\varphi}}_{1i}$ are identically independently distributed with error term $o_P(1)$. Thus, we could conclude that $\frac{\sum_{t_i=\mathcal{LC}}\widetilde{\boldsymbol{\alpha}}_1^\top\widetilde{\boldsymbol{\varphi}}_{1i}}{n}$ is bounded by $o_P(n^{-\frac{1}{2}})$. Finally, asymptotic properties of $\widetilde{\boldsymbol{\gamma}}_1$ have been guaranteed.

**Lemma 2.** $\widetilde{\boldsymbol{\varphi}}_2\xrightarrow{p}\bar{\boldsymbol{\varphi}}_2$, $\widetilde{\boldsymbol{\alpha}}_2\xrightarrow{p}\bar{\boldsymbol{\alpha}}_2$ *as* $n\to\infty$

**Theorem 3.**

$$\begin{aligned}&\sqrt{n}(\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_2-\boldsymbol{h}^\top\bar{\boldsymbol{\gamma}})\\&=\frac{\sum_{t_i=\mathcal{LC},\mathcal{LM}}\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_{1i}}{\sqrt{n}(1+\rho_1)}-\frac{\sum_{t_i=\mathcal{LC}}\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_{1i}}{\sqrt{n}}+\boldsymbol{h}^\top\boldsymbol{H}_{\bar{\boldsymbol{\gamma}}}^{-1}\frac{\sum_{t_i=C}\Phi_i}{\sqrt{n}}\\&+\frac{\sum_{t_i=\mathcal{LC},\mathcal{UC}}\bar{\boldsymbol{\alpha}}_2^\top\bar{\boldsymbol{\varphi}}_{2i}}{\sqrt{n}(1+\rho_2)}-\frac{\sum_{t_i=\mathcal{LC}}\bar{\boldsymbol{\alpha}}_2^\top\bar{\boldsymbol{\varphi}}_{2i}}{\sqrt{n}}+o_P(1)\end{aligned}$$

*In particular, the sequence* $\sqrt{n}(\widetilde{\boldsymbol{\gamma}}_2-\bar{\boldsymbol{\gamma}})$ *is asymptotically normal with mean zero and covariance matrix* $\mathbb{E}[(\boldsymbol{h}^\top\boldsymbol{H}_{\bar{\boldsymbol{\gamma}}}^{-1}\Phi-\frac{\rho_1}{1+\rho_1}\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1-\bar{\boldsymbol{\alpha}}_2^\top\bar{\boldsymbol{\varphi}}_2)^{\otimes 2}]+\frac{\rho_1}{(1+\rho_1)^2}\mathbb{E}[(\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^{\otimes 2}]+\frac{1}{1+\rho_2}\mathbb{E}[(\bar{\boldsymbol{\alpha}}_2^\top\bar{\boldsymbol{\varphi}}_2)^{\otimes 2}]$

The error term of $\widetilde{\boldsymbol{\gamma}}_2$ comprises of two parts, error of $\widetilde{\boldsymbol{\gamma}}_1$ and error of $\widetilde{\boldsymbol{\gamma}}_2$ to $\widetilde{\boldsymbol{\gamma}}_1$. We have shown in theorem 2 that error of $\widetilde{\boldsymbol{\gamma}}_1$ is $o_P(n^{-\frac{1}{2}})$. Similarly, we could conclude error of $\widetilde{\boldsymbol{\gamma}}_2$ to $\widetilde{\boldsymbol{\gamma}}_1$ is also $o_P(n^{-\frac{1}{2}})$ since $\widetilde{\boldsymbol{\alpha}}_2^\top\widetilde{\boldsymbol{\varphi}}_{2i}$ are i.i.d and $o_P(1)$. Thus, through CLT we could have two asymptotic normal distributions with the same means and error rates. In one more step, we could state the error rate of ($\widetilde{\boldsymbol{\gamma}}_2$ - $\bar{\boldsymbol{\gamma}}$) is $o_P(n^{-\frac{1}{2}})$ by subtracting both.

# 5 Efficiency of Asymptotic Variance

We have shown before that our proposed two estimators, Sip and Dop, are asymptotic normal with error rate $o_P(n^{-\frac{1}{2}})$ and we also derive the asymptotic variance of two estimators. Then we need to prove the asymptotic efficiency of our proposed estimator. Recall the expression of $\widetilde{\boldsymbol{\gamma}}_1$:

$$\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_1 = \boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}} - \frac{1}{n}\sum_{t_i=\mathcal{LC}}\widetilde{\boldsymbol{\alpha}}_1^\top\widetilde{\boldsymbol{\varphi}}_{1i} + \frac{1}{(1+\rho_1)n}\sum_{t_i=\mathcal{LC},\mathcal{LM}}\widetilde{\boldsymbol{\alpha}}_1^\top\widetilde{\boldsymbol{\varphi}}_{1i}$$

We should notice that $\widetilde{\boldsymbol{\alpha}}_1$ is the coefficient of the linear regression between score functions $(\boldsymbol{h}^\top\widetilde{\boldsymbol{S}}_i)$ and projection of score functions $(\widetilde{\boldsymbol{\varphi}}_{1i})$. The orthogonality between linear residual $(\boldsymbol{h}^\top\widetilde{\boldsymbol{S}}_i - \widetilde{\boldsymbol{\alpha}}_1^\top\widetilde{\boldsymbol{\varphi}}_{1i})$ and predictors $(\widetilde{\boldsymbol{\varphi}}_{1i})$ make great convenience to calculations of variance that we have the simple, comparable, closed form of our proposed estimators. Thus,

$$\begin{aligned}\mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_1) &= \mathbb{E}(\boldsymbol{h}^\top\bar{\boldsymbol{S}} - \bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^2 + \frac{\mathbb{E}(\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^2}{(1+\rho_1)}\\ \text{while}\quad \mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}) &= \mathbb{E}(\boldsymbol{h}^\top\boldsymbol{S} - \bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^2 + \mathbb{E}(\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^2\end{aligned}$$

The relative efficiency of Sip to LC-only (RE1) is defined as

$$\frac{\mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_1)}{\mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}})} = 1 - \frac{\rho_1}{1+\rho_1}\frac{1}{1+\frac{\mathbb{E}(\boldsymbol{h}^\top\bar{\boldsymbol{S}}-\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^2}{\mathbb{E}(\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^2}}$$

It is easy to see RE1 decreases when $\rho_1$ increases, which is consistent with the real situation since when we fix the sample size of labeled complete data as n, larger $\rho_1$ represents we have more labeled incomplete data whose potential distribution is the same as those of labeled complete data. Naturally, the proposed estimator would outperform the LC-only estimator. Besides, RE1 decreases when $\frac{\mathbb{E}(\boldsymbol{h}^\top\bar{\boldsymbol{S}}-\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^2}{\mathbb{E}(\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^2}$ decreases. Remember $\bar{\boldsymbol{\alpha}}_1$ coming from the linear regression, which indicates $(\boldsymbol{h}^\top\bar{\boldsymbol{S}} - \bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)$ and $\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1$ are orthogonal with each others. Thus, smaller $\frac{\mathbb{E}(\boldsymbol{h}^\top\bar{\boldsymbol{S}}-\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^2}{\mathbb{E}(\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^2}$ refers to larger $\frac{\mathbb{E}(\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^2}{\mathbb{E}(\boldsymbol{h}^\top\bar{\boldsymbol{S}})^2}$. Thus, when partial projections of score functions could explain more of original score functions, the $\mathcal{LM}$ data could do more favors to learn the coefficients.

Recall the expression of $\widetilde{\boldsymbol{\gamma}}_2$:

$$\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_2 = \boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_1 - \frac{1}{n}\sum_{t_i=\mathcal{LC}}\widetilde{\boldsymbol{\alpha}}_2^\top\widetilde{\boldsymbol{\varphi}}_{2i} + \frac{1}{(1+\rho_2)n}\sum_{t_i=\mathcal{LC},\mathcal{UC}}\widetilde{\boldsymbol{\alpha}}_2^\top\widetilde{\boldsymbol{\varphi}}_{2i}$$

Define $\mathrm{R}(\widetilde{\boldsymbol{\varphi}}_1) = \boldsymbol{h}^\top\widetilde{\boldsymbol{S}} - \frac{\rho_1}{1+\rho_1}\widetilde{\boldsymbol{\alpha}}_1^\top\widetilde{\boldsymbol{\varphi}}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y)$, which is the term related to $\mathcal{LC}$ in (3). We hope the second projection can capture as much information as possible. Thus apply regression between $\mathrm{R}(\widetilde{\boldsymbol{\varphi}}_1)$ and $\widetilde{\boldsymbol{\varphi}}_2$. In step 3, we apply linear regression between $\mathrm{R}(\widetilde{\boldsymbol{\varphi}}_{1i})$ and $\widetilde{\boldsymbol{\varphi}}_{2i}$. Thus, we can easily get the expression of variance:

$$\begin{aligned}\mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_2) &= \mathbb{E}(\mathrm{R}(\bar{\boldsymbol{\varphi}}_1) - \bar{\boldsymbol{\alpha}}_2^\top\bar{\boldsymbol{\varphi}}_2)^2 + \frac{\mathbb{E}(\bar{\boldsymbol{\alpha}}_2^\top\bar{\boldsymbol{\varphi}}_2)^2}{(1+\rho_2)} + \frac{\rho_1\mathbb{E}(\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^2}{(1+\rho_1)^2}\\ \text{while}\quad \mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_1) &= \mathbb{E}(\mathrm{R}(\bar{\boldsymbol{\varphi}}_1) - \bar{\boldsymbol{\alpha}}_2^\top\bar{\boldsymbol{\varphi}}_2)^2 + \mathbb{E}(\bar{\boldsymbol{\alpha}}_2^\top\bar{\boldsymbol{\varphi}}_2)^2 + \frac{\rho_1\mathbb{E}(\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^2}{(1+\rho_1)^2}\end{aligned}$$

Similarly, we define the relative efficiency of Dop to Sip (RE2) as:

$$\frac{\mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_2)}{\mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_1)} = 1 - \frac{\rho_2}{1+\rho_2}\frac{1}{1+\frac{\mathbb{E}(\mathrm{R}(\bar{\boldsymbol{\varphi}}_1)-\bar{\boldsymbol{\alpha}}_2^\top\bar{\boldsymbol{\varphi}}_2)^2}{\mathbb{E}(\bar{\boldsymbol{\alpha}}_2^\top\bar{\boldsymbol{\varphi}}_2)^2}+\frac{\rho_1\mathbb{E}(\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^2}{(1+\rho_1)^2\mathbb{E}(\bar{\boldsymbol{\alpha}}_2^\top\bar{\boldsymbol{\varphi}}_2)^2}}$$

It is easy to see RE2 decreases when $\rho_2$ increases since we have a larger sample size. However, the dependency of RE2 on $\rho_1$ is different. $\frac{\rho_1}{(1+\rho_1)^2}$ increases when $\rho_1 : 0 \to 1$ and decreases when $\rho_1 : 1 \to +\infty$. Since we want a smaller RE2, we hope a smaller $\frac{\rho_1}{(1+\rho_1)^2}$, which means that we hope $\rho_1 \to 0$ when $\rho_1 \leq 1$ and the larger $\rho_1$ the better when $\rho_1 \geq 1$. And RE2 decreases when $\frac{\mathbb{E}(\mathrm{R}(\bar{\boldsymbol{\varphi}}_1)-\bar{\boldsymbol{\alpha}}_2^\top\bar{\boldsymbol{\varphi}}_2)^2}{\mathbb{E}(\bar{\boldsymbol{\alpha}}_2^\top\bar{\boldsymbol{\varphi}}_2)^2}$ and $\frac{\mathbb{E}(\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1)^2}{\mathbb{E}(\bar{\boldsymbol{\alpha}}_2^\top\bar{\boldsymbol{\varphi}}_2)^2}$ decreases separately. Remembering that $\bar{\boldsymbol{\varphi}}_2$ is the partial projection of $\bar{\boldsymbol{\varphi}}_1$, smaller values meaning that $\bar{\boldsymbol{\varphi}}_2$ could explain more about $\bar{\boldsymbol{\varphi}}_1$ and $\mathrm{R}(\bar{\boldsymbol{\varphi}}_1) = \boldsymbol{h}^\top\bar{\boldsymbol{S}} - \frac{\rho_1}{1+\rho_1}\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1$. However, $\bar{\boldsymbol{\varphi}}_1$ and $\mathrm{R}(\bar{\boldsymbol{\varphi}}_1) = \boldsymbol{h}^\top\bar{\boldsymbol{S}} - \frac{\rho_1}{1+\rho_1}\bar{\boldsymbol{\alpha}}_1^\top\bar{\boldsymbol{\varphi}}_1$ are somehow orthogonal with each other, that's could be one potential explanation of not so good performance of Dop compared to Sip. But, we could introduce more $\mathcal{LM}$ and $\mathcal{UC}$ (larger $\rho_1\rho_2$) to overcome the conflicts we mentioned before.

In the above section, we analyze the general asymptotic variance efficiency of our proposed SiP and DoP estimators. Remember we apply the machine learning model when calculating the projection of the score function $\widetilde{\boldsymbol{\varphi}}_1$. If we further require the machine learning model to be consistent with the true model, which is standard when utilizing some common methods such as random forest and non-parametric model, we could have the simplified formula of relative efficiency. Suppose the machine learning model is consistent and $\boldsymbol{\varphi}_1 = \mathbb{E}[\boldsymbol{S} \mid \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y]$, we believe the estimation of $\boldsymbol{\varphi}_1$ is "correctly specified". Thus, the coefficient $\widetilde{\boldsymbol{\alpha}}_1$ of the linear regression should be exactly equal to $\boldsymbol{h}$. Thus, $\frac{\mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_1)}{\mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}})}$ can be simplified as $1 - \frac{\rho_1}{\rho_1+1}\frac{\mathrm{Var}(\mathbb{E}(\boldsymbol{h}^\top\bar{\boldsymbol{S}}|\boldsymbol{X}_{\mathcal{P}_{m^c}},Y))}{\mathrm{Var}(\boldsymbol{h}^\top\bar{\boldsymbol{S}})}$.

Similarly, we suppose $\boldsymbol{\varphi}_2 = \mathbb{E}[\boldsymbol{\varphi}_1 \mid \boldsymbol{X}]$. When the above model is "correctly specified", the coefficient $\widetilde{\boldsymbol{\alpha}}_2$ of the linear regression should be exactly equal to $\frac{1}{1+\rho_1}\boldsymbol{h}$.And we compare $\widetilde{\boldsymbol{\gamma}}_2$ with $\widetilde{\boldsymbol{\gamma}}_1$:

$$\mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_2) = \mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_1) - \frac{\rho_2}{(1+\rho_2)}\mathrm{Var}(\mathbb{E}(\boldsymbol{h}^\top\bar{\boldsymbol{S}} - \mathbb{E}(\frac{\rho_1}{1+\rho_1}\boldsymbol{h}^\top\bar{\boldsymbol{S}} \mid \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y)) \mid \boldsymbol{X})$$

$\frac{\mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_2)}{\mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}})}$ has the further reduction term $\frac{\rho_2}{1+\rho_2}\frac{\mathrm{Var}(\mathbb{E}(\boldsymbol{h}^\top\bar{\boldsymbol{S}}-\mathbb{E}(\frac{\rho_1}{1+\rho_1}\boldsymbol{h}^\top\bar{\boldsymbol{S}}|\boldsymbol{X}_{\mathcal{P}_{m^c}},Y))|\boldsymbol{X})}{\mathrm{Var}(\boldsymbol{h}^\top\bar{\boldsymbol{S}})}$ compared to $\frac{\mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}}_1)}{\mathrm{Var}(\sqrt{n}\boldsymbol{h}^\top\widetilde{\boldsymbol{\gamma}})}$. We could also conclude from the above equation that the better estimation of projection for score function, the better our proposed SiP and DoP could be.

# 6 Simulation Study

In this section, we present a comprehensive simulation study designed to evaluate and compare the performance of our proposed estimator with a selection of benchmark estimators. The selected benchmarks include: an estimator from semi-supervised learning using complete data and unlabeled data, an estimator obtained post-Multiple Imputation by Chained Equations (MICE) for handling incomplete data, and the proposed estimator that only utilize either complete and incomplete data without employing unlabeled data. Notably, the

robustness of our method was investigated in both correctly specified and mis-specified model scenarios to prove its efficiency. To enhance the credibility of our simulation study, we also explored scenarios involving different imputation methods such as Support Vector Machine (SVM) and Lasso regression. This was important to ensure that the performance of our proposed method remained robust, independent of the imputation approach employed for handling missing data.

## 6.1 Data Generation

For the data generation phase of our simulation, we generated five dimensions of covariates, $p = 5$. The data for the first three dimensions were generated separately from uniform distribution $\boldsymbol{U}(-1, 1)$. The data for the fourth dimension $X_4$ was generated from $N(0.8X_1, \sigma)$, and the $X_5$ was generated from $N(0.1X_1, \sigma)$, where $\sigma = 0.6$. Here the $X_4$ and $X_5$ were assumed to be missing in $\mathcal{X}_{\mathcal{LM}}$. The imputation method we used for these two dimensions was support vector machine. For the generation of dependent variable, which was a binary form, we used the correctly specified model as follow. Further, we used misspecified model to generate outcome and tested the robustness of our method in the Appendix.

$$logit(P(Y = 1|\boldsymbol{X})) = \boldsymbol{X}\boldsymbol{\gamma}$$

True $\boldsymbol{\gamma}$ was defined as $(0, 0.4, -0.4, -0.4, 0.8, -0.8)$. Firstly, we got the $P(Y_i = 1|\boldsymbol{X_i})$ from the correctly specified model above for each observation, then we generated $Y_i$ from $Bernoulli(P(Y_i = 1|\boldsymbol{X_i}))$.

We also changed the $N_{\mathcal{LC}}$ to see the influence that the ratio of incomplete sample size and complete sample size has on the performance of our estimator. The number of complete data is fixed at $N_{\mathcal{LC}} = 300$. The number of unlabeled data is fixed at $N_{\mathcal{UC}} = 2700$. We changed the number of incomplete data from 600 to 1800, $N_{\mathcal{LM}} = 600, 900, 1200, 1500, 1800$. All the simulation settings were repeated 500 times to get the results.

## 6.2 Performance Metrics

Four performance metrics were evaluated for the estimators. They are bias, relative efficiency, prediction mean squared error, and coverage rate.

As shown in Table 1, we can see that Mice had much larger bias on missing dimensions so that it was not considered for further performance evaluation. Our proposed Sip and Dop methods had no bias.

From Table 2, we can see that the average relative efficiency across dimensions of Sip and Dop are greater than one, which means they reduced the variance of estimator a lot. Besides, Dop estimator had higher relative efficiency than the Sip. What's more, with the increasing of incomplete sample size, the relative efficiency of both Sip estimator and Dop estimator

Table 1: Bias for different methods

| Method | Dimensions | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Naive | 0.0036 | 0.027 | -0.016 | -0.028 | -0.020 | -0.029 |
| Sip | 0.0042 | -0.0068 | -0.0026 | -0.0078 | 0.014 | -0.024 |
| Dop | 0.0032 | -0.023 | -0.0051 | -0.012 | -0.0097 | -0.003 |
| SSL | 0.0038 | 0.027 | -0.015 | -0.027 | 0.020 | -0.028 |
| Mice | 0.0018 | 0.21 | 0.028 | 0.024 | -0.41 | 0.387 |

increased. Moreover, we also tried linear imputation and lasso imputation, and the relative efficiency of them showed difference smaller than 5%.

Table 2: Relative Efficiency

| Method | Incomplete Sample Size | | | | |
|---|---|---|---|---|---|
| | 600 | 900 | 1200 | 1500 | 1800 |
| Naive | 1 | 1 | 1 | 1 | 1 |
| Sip | 1.87 | 2.12 | 2.36 | 2.47 | 2.81 |
| Dop | 2.02 | 2.35 | 2.69 | 3.09 | 3.45 |
| SSL | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

For dimension wise relative efficiency of Sip and Dop, we could see from Figure 2 that the REs for Dop were higher than those of Sip in complete dimensions. But in the missing dimension, the relative efficiency of Sip and Dop might not be that obvious.

We also made a figure about prediction mean squared error. In Figure 3, we can see that the prediction mean squared errors of dop are lower than those of naive method and semi-supervised learning method. It could reduce the MSE of Naive estimator by 74.5%, 43.9%, 74.4% and 74.8% in the first four dimensions. Besides, as the incomplete sample size went up, the MSE of proposed method went down. We also showed our propose methods could lower prediction mean square error even when the outcome model was incorrectly specified, which was shown in Appendix.

The 95% confidence interval for proposed estimator also had good coverage rate, as shown in Table 3.

Table 3: Coverage Rate

| Method | Incomplete Sample Size | | | | |
|---|---|---|---|---|---|
| | 600 | 900 | 1200 | 1500 | 1800 |
| Dop | 94.67% | 94.40% | 94.17% | 95.30% | 95.03% |

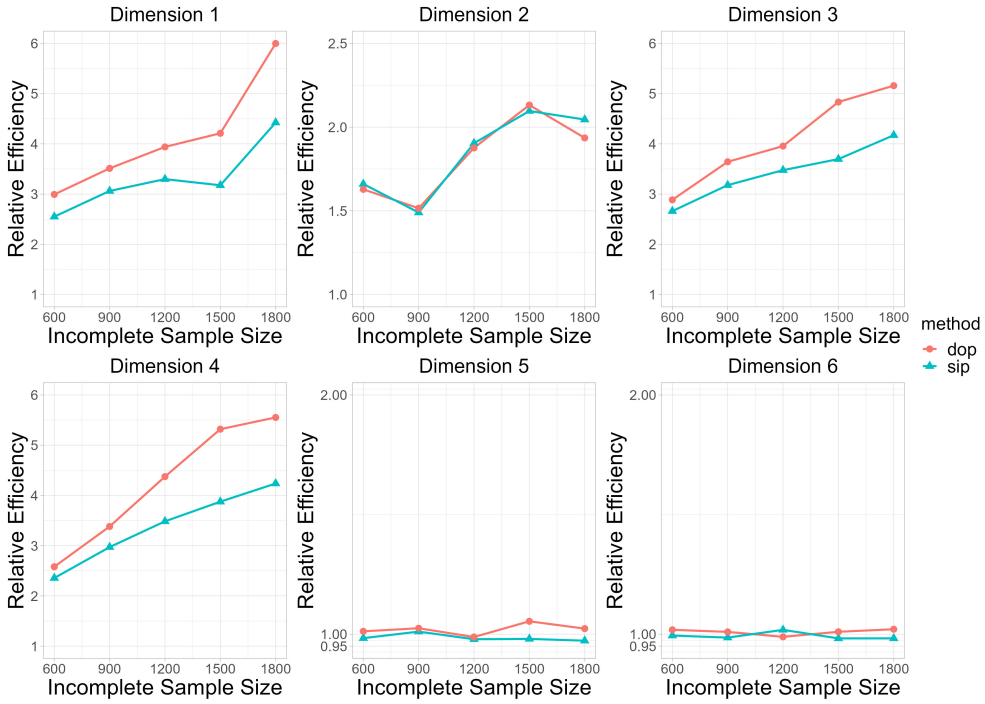


Figure 2: Relative Efficiency of Sip and Dop

# 7 Real Example

MIMIC-III (Medical Information Mart for Intensive Care III) is a large, freely available database comprising de-identified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. more. The database includes information such as demographics, vital sign measurements made at the bedside (1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of the hospital). However, the MIMIC III data has much missing. We hope to apply our proposed method to build a better estimation.

**preliminary** The first thing we need to make sure is how to determine the response of data. Gehrman constructed a clinical trial to compare whether frequent patients (defined as $\geq 3$ ICU visits within 365 days) and non-frequent patients have different distributions of phenotypes. In this study, they use lab measurements and diagnostic results from multiple physicians as the golden labeled data (response we want) for a total of 1045 patients. The golden labels include Metastatic Cancer, Heart Disease, Lung Cancer, Chronic Neurologic, Dystrophies, Chronic Pain, Alcohol Abuse, Substance Abuse, Obesity, Psychiatric disorders, and Depression.(1)

We then construct our potential data by including demographic features (age, gender, ethnicity) as $\boldsymbol{X}_{\mathcal{P}_{m^c}}$. These features are easy to obtain and are known to be related to the diseases mentioned above. **cite**. We also include some clinical variables such as the main surrogate which represents the count of the main ICD records, and the healthcare utilization which is defined as the days of admission over the total count of ICD codes.

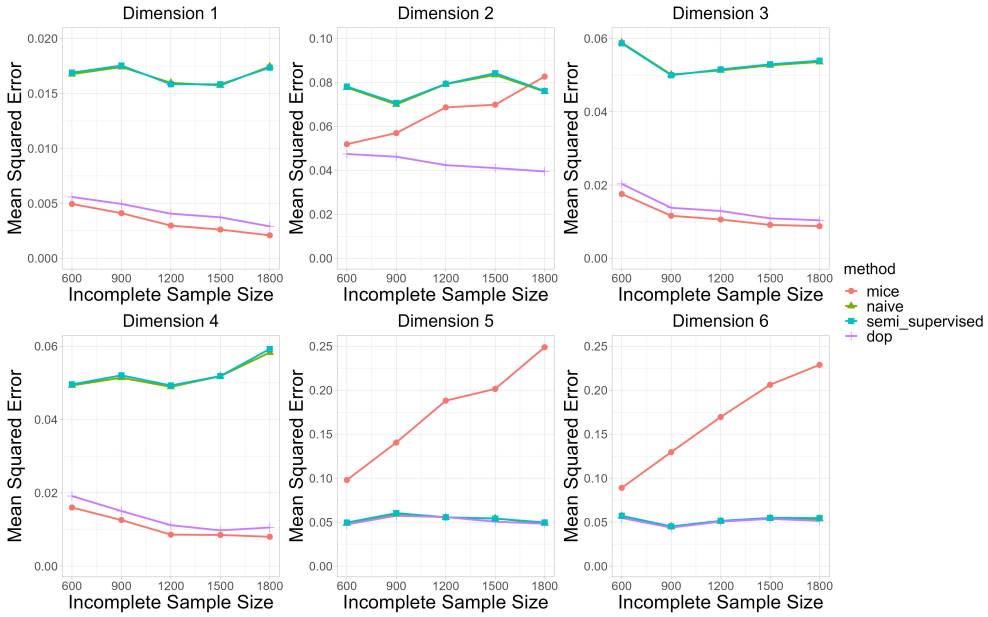


Figure 3: MSE Comparison between methods and imcomplete sample sizes

We then consider the lab values as the potential missing for the regression model. We require the added lab values to be both correlated with the main diseases and valid to improve the simple logistic regression after adding. So we first search all the correlated variables through **KESER**. We further determine the validity of specific lab data by comparing the AUC with and without the addition of this lab data.

We finally chose hormones in Psychiatric disease and microcytes in lung dieases as potentially missing. This lab value is chosen with the following rules: 1. The missing rate of ovalocytes should be almost the same across either labeled or unlabeled data. 2. The AUC of simple logistic regression should be larger after including this feature compared to the case when only regressing golden label data to main surrogates and health utilization. 3. The missing rate in two groups (main surrogate $= 0$ and main surrogate $> 0$) should be practically equal. **The plot of AUC is shown in the appendix.**

Table 4 shows the coefficient value of LC-only estimator, our proposed Sip and Dop estimators, MICE estimator, and a semi-supervised learning estimator. All the coefficients are the average of 10 runs. We might conclude the coefficients are unbiased.

Table 4: Estimated parameters

| Features | Naive | Sip | Dop | Mice | SSL |
|---|---|---|---|---|---|
| (Intercept) | -4.17 | -5.00 | -5.01 | -5.07 | -4.46 |
| psychiatric count | 2.51 | 2.41 | 2.40 | 2.35 | 2.51 |
| lab | 0.33 | 0.34 | 0.30 | 0.31 | 0.33 |
| heal utilization | 0.26 | 0.55 | 0.53 | 0.46 | 0.38 |

As the complete sample size is greater than the incomplete sample size, we randomly sample the complete data equal to and smaller than the incomplete sample size. We do it

10 times and average the result. Table 5 shows the relative efficiency of the variance for different $\frac{\mathcal{LC}}{\mathcal{LM}}$ ratios compared to the LC-only estimator.

Table 5: Relative Efficiency

| Feature | ratio = 1 | | ratio = 2 | | ratio = 3 | |
|---|---|---|---|---|---|---|
| | Sip | Dop | Sip | Dop | Sip | Dop |
| (Intercept) | 1.187 | 1.188 | 1.299 | 1.308 | 1.399 | 1.425 |
| psychiatric count | 1.811 | 1.833 | 2.327 | 2.357 | 2.812 | 2.882 |
| lab | 1.004 | 1.022 | 1.018 | 1.061 | 1.036 | 1.097 |
| health utilization | 1.899 | 1.904 | 2.739 | 2.817 | 3.336 | 3.530 |

When screening the useful lab values, we find some specific lab values have similar missing tendencies among the population. For example, more than 95 % of the individuals have microcyte, macrocyte, and red cell count at the same time. This motivates us to combine the lab value with correlated clinical meaning and similar missing structure. We first select five target lab values using the above procedure and criterion. Then we choose all the lab values that overlap more than 95 % condition on the target lab and impute the part in which the target lab is observed but selected labs are missed by the mean value. We further define the lab score as the prediction value of main surrogates over the imputed lab and choose the suitable golden label which increases the AUC of logistic regression after including the lab score compared to the case when only regressing golden label data to main surrogates. The AUC and combined labs are provided in the Appendix.

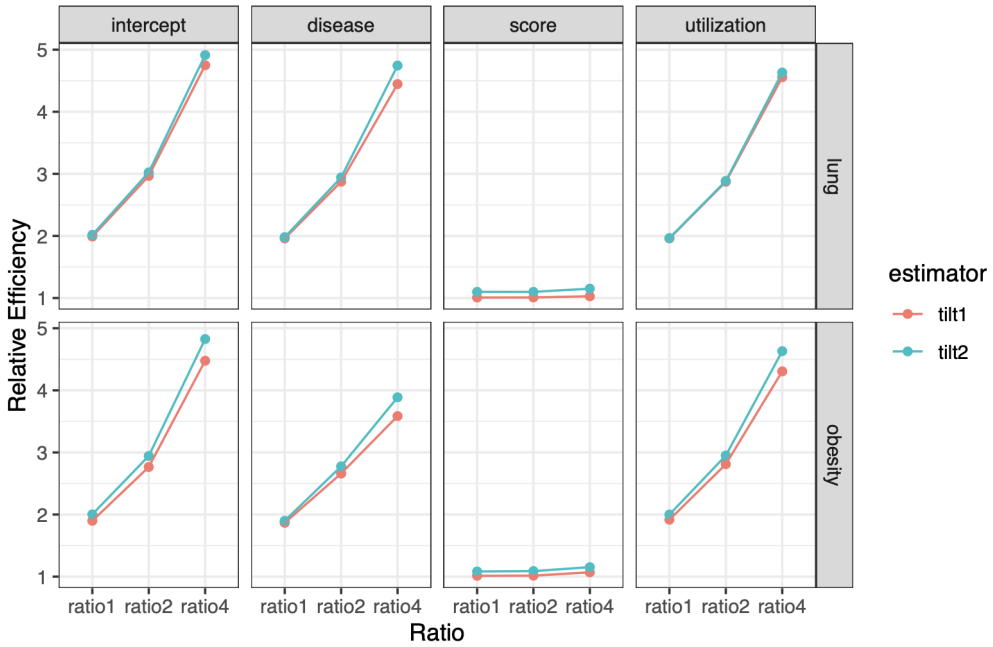


Figure 4: Relative Efficiency For Obesity and Lung Diseases

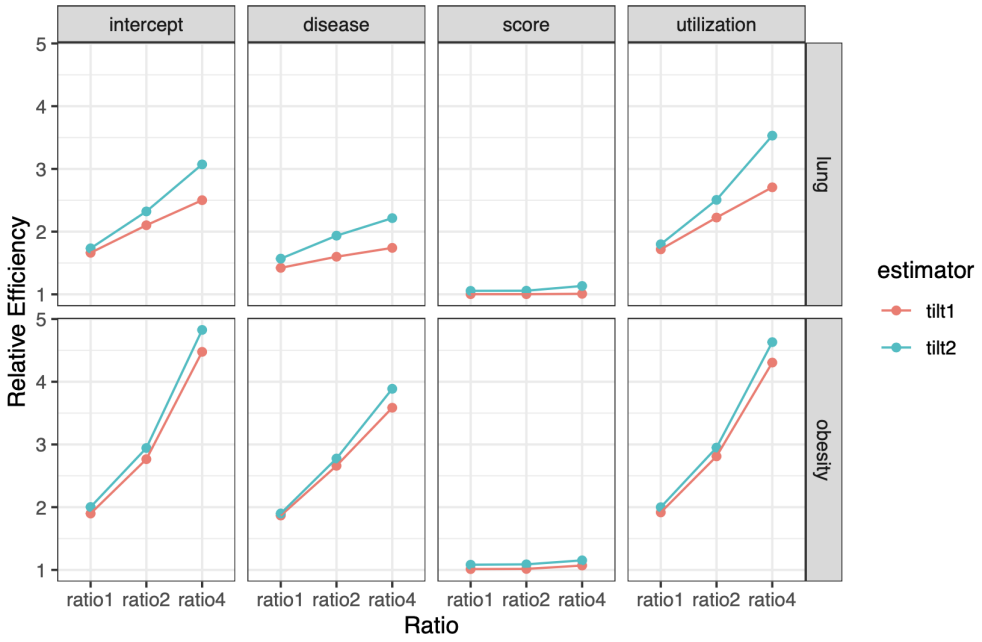


Figure 5: Relative Efficiency For Obesity and Lung Diseases

In figure 4, we show the relative efficiency of our proposed SiP and DoP estimators compared to the LC-only estimator. The disease is the corresponding main surrogate count of obesity and advanced lung disease (COPD). The target lab for COPD is Ovalocytes and the score for COPD is the prediction value over Ovalocytes, Red Blood Cells, Anisocytosis, Hypochromia, Macrocytes, Microcytes, and Poikilocytosis. The target lab for obesity is Alveolar-arterial Gradient and the score for obesity is the prediction value over Alveolar-arterial Gradient, Red Blood Cells, Base Excess, Oxygen, pH, pO2, and PT. The ratio represents the sample size ratio of $\mathcal{LM}$ data over the $\mathcal{LC}$ data, i.e., $\rho_1$ in the asymptotic section. We do not include the variation of $\rho_2$ since we always assume that $\mathcal{UC}$ data is much more than $\mathcal{LC}$ data. The relative efficiency of other available scores is provided in the Appendix

**lung 1 the same to obesity**

The total performance of our proposed SiP and DoP estimators is better than the LC-only estimator since all relative efficiency is greater than 1. The increase of $\rho_1$ could improve the performance of the algorithm, which is consistent with the conclusion we have made in the above asymptotic section. The DoP always outperforms the SiP. The result is natural since we include more $\mathcal{UC}$, even though the improvement is not so obvious.

We also show in table 6 that coefficients are unbiased among the LC-only estimator, SiP, and DoP. For Obesity, the intercept of the LC-only estimator is -5.304 with a standard error equal to 1.489, while the coefficients are -4.057 and -3.985, and standard errors are 1.080 and 1.052 for Sip and Dop. Even if the estimation of the intercept might not be so close, we still believe they are equal since the confidence intervals of these three have large overlaps. Remember we randomly draw a subsample from the $\mathcal{LC}$ to increase while keeping $\mathcal{LM}$ unchanged when we try to increase $\rho_1$. However, there is a potential problem. Most

Table 6: Relative Efficiency

| Ratio = 1 | | LC-only | Sip | Dop |
|---|---|---|---|---|
| Lung2 | (Intercept) | -3.382±1.100 | -3.985±0.780 | -4.244±0.774 |
| | Disease | 2.225±0.414 | 2.134±0.296 | 2.170±0.294 |
| | Score | 0.424±0.288 | 0.443±0.287 | 0.954±0.274 |
| | Utilization | 0.220±0.278 | 0.362±0.198 | 0.405±0.197 |
| Obesity | (Intercept) | -5.304±1.489 | -4.057±1.080 | -3.985±1.052 |
| | Disease | 3.251±0.442 | 3.742±0.324 | 3.662±0.321 |
| | Score | 0.384±0.211 | 0.373±0.210 | 0.519±0.203 |
| | Utilization | 0.543±0.350 | 0.296±0.253 | 0.210±0.248 |
| Lung1 | (Intercept) | -3.587±0.943 | -4.500±0.731 | -3.907±0.717 |
| | Disease | 1.128±0.457 | 1.070 ±0.383 | 1.310±0.365 |
| | Score | 1.091±0.193 | 1.087±0.193 | 1.108±0.188 |
| | Utilization | 0.261±0.226 | 0.431±0.173 | 0.290±0.169 |

responses of the subsample tend to be the same (either 0 or 1) when we extract the smaller sample from $\mathcal{LC}$. The coefficient seems to be unstable when the ratio is too large for MIMIC III data. So we only show the coefficients when ratio equals to 1 for some target labs, the rest results are provided in the Appendix.

# 8 Discussion

# References

[1] Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, Foote J Jr, Moseley ET, Grant DW, Tyler PD, Celi LA. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS One. 2018 Feb 15;13(2):e0192360. doi: 10.1371/journal.pone.0192360. PMID: 29447188; PMCID: PMC5813927.

# Appendix

# A equivalent of the cross fitting estimator

For the K-fold cross-fitting method, we first suppose the complete data is randomly split into $K$ different samples with roughly equal sample size, denoted by $\{C^{(1)}, C^{(2)}, ..., C^{(K)}\}$, where $|C^{(k)}| \approx \frac{n}{k}$ and $C^{(-k)} = \{C^{(j)} : j \neq k\}$ for $k \in [1 : K]$. For each $k$, apply machine learning methods such as random forest or kernel smooth to regress $\boldsymbol{X}_{\mathcal{P}_m} \sim \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y$ with respective to nuisance estimator $\bar{\boldsymbol{\eta}}$, denoted as $\widetilde{\boldsymbol{\eta}}^{(-k)}$ based on the data $C^{(-k)}$. Thus, on the fold $C^{(-k)}$:

$$\begin{aligned}
&\boldsymbol{X}_{\mathcal{P}_m} = f(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y; \bar{\boldsymbol{\eta}}) + \epsilon, \text{ where } f(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y; \widetilde{\boldsymbol{\eta}}) \text{ is estimation} \\
&\boldsymbol{X}_{\mathcal{P}_m} = f^k(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y; \bar{\boldsymbol{\eta}}^{-k}) + \epsilon, \text{ where } f^k(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y; \widetilde{\boldsymbol{\eta}}^{-k}) \text{ is estimation based on fold } C^{(-k)}
\end{aligned}$$

For the evaluation process, we estimate $\boldsymbol{X}_{i\mathcal{P}_m} \mid \boldsymbol{X}_{i\mathcal{M}^c}, Y_i$ based on $\widetilde{\boldsymbol{\eta}}^{(-k)}$ for each $i \in IC$ and the results are recorded as $f^k(\boldsymbol{X}_{i\mathcal{M}^c}, Yi; \widetilde{\boldsymbol{\eta}}^{-k})$ for $i \in IC$. The average K-fold cross-fitting estimation $\widetilde{\boldsymbol{X}}^{CF}_{i\mathcal{M}} \mid \boldsymbol{X}_{i\mathcal{M}^c}, Y_i = \frac{1}{K}\sum_{k=1}^{K} f^k(\boldsymbol{X}_{i\mathcal{M}^c}, Yi; \widetilde{\boldsymbol{\eta}}^{-k})$. First, machine learning methods like random forest and kernel smooth have nice consistency properties. Namely, we could denote $\widetilde{\boldsymbol{\eta}}^{(-k)}$ is consistent to $\bar{\boldsymbol{\eta}}^{(-k)}$ on the fold $C^{(-k)}$ in the complete data set. We also assume that the potential distribution of $\boldsymbol{X}_{i\mathcal{M}^c}, Yi$ are homogenous across the complete and incomplete data sets. Thus, $f^k(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y; \widetilde{\boldsymbol{\eta}}^{-k}) - f^k(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y; \bar{\boldsymbol{\eta}}^{-k})$ is bounded by $o_P(1)$ in the incomplete data set. It is not hard to see $\bar{\boldsymbol{\eta}}^{-k}$ converges to $\bar{\boldsymbol{\eta}}$ in probability when k < n. Thus we have the following convergence chain:

$$\frac{1}{K}\sum_{k=1}^{K} f^k(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y; \widetilde{\boldsymbol{\eta}}^{-k}) \xrightarrow{p} \frac{1}{K}\sum_{k=1}^{K} f^k(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y; \bar{\boldsymbol{\eta}}^{-k}) \xrightarrow{p} f(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y; \bar{\boldsymbol{\eta}})$$

Thus, if the cross-fitting estimator $\frac{1}{K}\sum_{k=1}^{K} f^k(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y; \widetilde{\boldsymbol{\eta}}^{-k})$ is consistent, the

$$\widetilde{\boldsymbol{\varphi}}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y) = \int \widetilde{\boldsymbol{S}}\widetilde{P}(\boldsymbol{X}_{\mathcal{P}_m} \mid \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y)\, d\boldsymbol{X}_{\mathcal{P}_m}$$

is also consistent.

# B Asymptotic of LC-only Estimator

The justification of the asymptotic properties is pretty standard. We first show the consistency and then show its asymptotic property. The naive model parameter, $\widetilde{\boldsymbol{\gamma}}$, is the solution to the following estimating equation:

$$\widetilde{\mathbf{U}}_{N_{\mathcal{LC}}}(\boldsymbol{\gamma}) = \frac{1}{N_{\mathcal{LC}}} \sum_{t_i = \mathcal{LC}} \boldsymbol{X}_i\{Y_i - g(\boldsymbol{\gamma}^\top \boldsymbol{X}_i)\} = \mathbf{0}$$

Although we might not have the closed form of expression for the numerical estimation, the deriving of this estimator comes is processed by maximizing the likelihood, which results in an MLE(Maximum Likelihood Estimation). By **XXXXX**, we could easily conclude the consistency of MLE. Under conditions 1-6, $\boldsymbol{X}_i$ belong to the compact space, and $g(\bar{\boldsymbol{\gamma}}^\top \boldsymbol{X}_i)$ is continuous and uniformly bounded for $\bar{\boldsymbol{\gamma}} \in \boldsymbol{\Gamma}$. From the **uniform law of large numbers (ULLN)**.[Pollard, 1990, Theorem8.2], $\boldsymbol{H}^{-1}\boldsymbol{X}_i\{Y_i - g(\widetilde{\boldsymbol{\gamma}}^\top \boldsymbol{X}_i)\}$ converge to $\mathbb{E}[\boldsymbol{H}^{-1}\boldsymbol{X}_i\{Y_i - g(\widetilde{\boldsymbol{\gamma}}^\top \boldsymbol{X}_i)\}]$ in probability as $N_{\mathcal{LC}} \to \infty$ and sup by $o_P(1)$. Furthermore, we consider the asymptotic normality, we apply Theorem **5.21 of Van der Vaart [2000]** to obtain the Taylor expansion. It then follows the general central limit theorem.

# C Asymptotic Property SiP Estimation

We begin by showing that $\widetilde{\boldsymbol{\gamma}}_1 \xrightarrow{p} \bar{\boldsymbol{\gamma}}$. We note that

$$\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_1 = \boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}} - \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i} + \frac{1}{N_{\mathcal{LC}} + N_{\mathcal{LM}}} \sum_{t_i=\mathcal{LC},\mathcal{LM}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}$$

We first note that $\widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}$ follows ULLN since $\widetilde{\boldsymbol{\varphi}}_1$ is continuously differentiable in $\widetilde{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\eta}}$ and uniformly bounded. Besides, the complete data set and the incomplete data set follow the same distribution. Thus, $\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}$ and $\frac{1}{N_{\mathcal{LC}}+N_{\mathcal{LM}}} \sum_{t_i=\mathcal{LC},\mathcal{LM}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}$ converge in probability to the same item and are canceled. The consistency of $\widetilde{\boldsymbol{\gamma}}_1$ for $\bar{\boldsymbol{\gamma}}$ then follows from the fact that $\widetilde{\boldsymbol{\gamma}}$ converge in probability to $\bar{\boldsymbol{\gamma}}$. To establish the asymptotic distribution of $\widetilde{\boldsymbol{\gamma}}_1$, we consider the summation of two independent mean zero normal distributions, $\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}$ and $-\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i} + \frac{1}{N_{\mathcal{LC}}+N_{\mathcal{LM}}} \sum_{t_i=\mathcal{LC},\mathcal{LM}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}$. Finally, we have

$$\begin{aligned}
&\sqrt{N_{\mathcal{LC}}}(\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_1 - \boldsymbol{h}^\top \bar{\boldsymbol{\gamma}}) \\
&= \frac{\sqrt{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC},\mathcal{LM}} \bar{\boldsymbol{\alpha}}_1^\top \bar{\boldsymbol{\varphi}}_{1i}}{N_{\mathcal{LC}} + N_{\mathcal{LM}}} - \frac{\sum_{t_i=\mathcal{LC}} \bar{\boldsymbol{\alpha}}_1^\top \bar{\boldsymbol{\varphi}}_{1i}}{\sqrt{N_{\mathcal{LC}}}} - \boldsymbol{h}^\top \boldsymbol{V}_{\bar{\boldsymbol{\gamma}}}^{-1} \frac{\sum_{t_i=C} \Phi_i}{\sqrt{N_{\mathcal{LC}}}} + o_P(1)
\end{aligned}$$

We also have $\sqrt{N_{\mathcal{LC}}}(\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_1 - \boldsymbol{h}^\top \bar{\boldsymbol{\gamma}}) \to N(0, \Sigma_{\widetilde{\gamma}_1})$ where $\Sigma_{\widetilde{\gamma}_1} = \frac{\mathbb{E}\left(\boldsymbol{h}^\top \bar{\boldsymbol{S}} - \bar{\boldsymbol{\alpha}}_1^\top \bar{\boldsymbol{\varphi}}_1\right)^2}{N_{\mathcal{LC}}} + \frac{\mathbb{E}\left(\bar{\boldsymbol{\alpha}}_1^\top \bar{\boldsymbol{\varphi}}_1\right)^2}{N_{\mathcal{LC}}+N_{IC}}$. We plug the expression of $\widetilde{\boldsymbol{\gamma}}$ into the expression of $\widetilde{\boldsymbol{\gamma}}_1$. Then we have

$$\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_1 = \boldsymbol{h}^\top \bar{\boldsymbol{\gamma}} + \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \boldsymbol{h}^\top \widetilde{\boldsymbol{S}}_i - \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i} + \frac{1}{N_{\mathcal{LC}} + N_{\mathcal{LM}}} \sum_{t_i=\mathcal{LC},\mathcal{LM}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}$$

And

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_1) &= \mathbb{E}[(\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \boldsymbol{h}^\top \widetilde{\boldsymbol{S}}_i - \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i} + \frac{1}{N_{\mathcal{LC}} + N_{\mathcal{LM}}} \sum_{t_i=\mathcal{LC},\mathcal{LM}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i})^2] \\
&= \mathbb{E}\{[\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} (\boldsymbol{h}^\top \widetilde{\boldsymbol{S}}_i - \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}) + \frac{1}{N_{\mathcal{LC}} + N_{\mathcal{LM}}} \sum_{t_i=\mathcal{LC},\mathcal{LM}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}]^2\}
\end{aligned}$$

Remember that $\widetilde{\boldsymbol{\alpha}}_1$ is the solution to linear regression between $\boldsymbol{h}^\top \widetilde{\boldsymbol{S}}_i$ and $\widetilde{\boldsymbol{\varphi}}_{1i}$. Thus, $\boldsymbol{h}^\top \widetilde{\boldsymbol{S}}_i - \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}$ should be orthogonal to $\widetilde{\boldsymbol{\varphi}}_{1i}$. All the individuals are independent. We have

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_1) &= \mathbb{E}\{[\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} (\boldsymbol{h}^\top \widetilde{\boldsymbol{S}}_i - \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}) + \frac{1}{N_{\mathcal{LC}} + N_{\mathcal{LM}}} \sum_{t_i=\mathcal{LC},\mathcal{LM}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}]^2\} \\
&= \mathbb{E}\{[\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} (\boldsymbol{h}^\top \widetilde{\boldsymbol{S}}_i - \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i})]^2\} + \mathbb{E}\{[\frac{1}{N_{\mathcal{LC}} + N_{\mathcal{LM}}} \sum_{t_i=\mathcal{LC},\mathcal{LM}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}]^2\} \\
&= \frac{\mathbb{E}\left(\boldsymbol{h}^\top \widetilde{\boldsymbol{S}} - \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_1\right)^2}{N_{\mathcal{LC}}} + \frac{\mathbb{E}\left(\widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_1\right)^2}{N_{\mathcal{LC}} + N_{IC}}
\end{aligned}$$

Similarly, the variance of $\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}$ can be estimated

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}) &= \mathbb{E}\{[\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} (\boldsymbol{h}^\top \widetilde{\boldsymbol{S}}_i - \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}) + \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}]^2\} \\
&= \mathbb{E}\{[\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} (\boldsymbol{h}^\top \widetilde{\boldsymbol{S}}_i - \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i})]^2\} + \mathbb{E}\{[\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}]^2\} \\
&= \frac{\mathbb{E}\left(\boldsymbol{h}^\top \widetilde{\boldsymbol{S}} - \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_1\right)^2}{N_{\mathcal{LC}}} + \frac{\mathbb{E}\left(\widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_1\right)^2}{N_{\mathcal{LC}}}
\end{aligned}$$

We can also see our proposed estimator has guaranteed efficiency since $\frac{\mathbb{E}\left(\widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_1\right)^2}{N_{\mathcal{LC}}} > \frac{\mathbb{E}\left(\widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_1\right)^2}{N_{\mathcal{LC}}+N_{\mathcal{LM}}}$. And the variance ratio would depend on $\mathbb{E}\left(\widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_1\right)^2$ and sample size $N_{\mathcal{LC}}, N_{\mathcal{LM}}$.

# D Asymptotic Property of DoP Estimation

We begin by showing that $\widetilde{\boldsymbol{\gamma}}_2 \xrightarrow{p} \bar{\boldsymbol{\gamma}}$. We note that

$$\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_2 = \boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_1 - \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i} + \frac{1}{N_{\mathcal{LC}}+N_{\mathcal{UC}}} \sum_{t_i=\mathcal{LC},\mathcal{UC}} \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i}$$

We first note that $\widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i}$ follows ULLN since $\widetilde{\boldsymbol{\varphi}}_2$ is continuously differentiable in $\widetilde{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\eta}}$ and uniformly bounded. Besides, the complete data set and the incomplete data set follow the same distribution. Thus, $\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i}$ and $\frac{1}{N_{\mathcal{LC}}+N_{\mathcal{UC}}} \sum_{t_i=\mathcal{LC},\mathcal{UC}} \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i}$ converge in probability to the same item and are canceled. The consistency of $\widetilde{\boldsymbol{\gamma}}_2$ for $\bar{\boldsymbol{\gamma}}$ then follows from the fact that $\widetilde{\boldsymbol{\gamma}}_1$ converge in probability to $\bar{\boldsymbol{\gamma}}$. To establish the asymptotic distribution of $\widetilde{\boldsymbol{\gamma}}_2$, we consider the summation of two independent mean zero normal distributions, $\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_1$ and $-\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i} + \frac{1}{N_{\mathcal{LC}}+N_{\mathcal{UC}}} \sum_{t_i=\mathcal{LC},\mathcal{UC}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}$. Finally, we have

$$\begin{aligned}
&\sqrt{N_{\mathcal{LC}}}(\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_2 - \boldsymbol{h}^\top \bar{\boldsymbol{\gamma}}) \\
&= \frac{\sqrt{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC},\mathcal{LM}} \bar{\boldsymbol{\alpha}}_1^\top \bar{\boldsymbol{\varphi}}_{1i}}{N_{\mathcal{LC}}+N_{\mathcal{LM}}} - \frac{\sum_{t_i=\mathcal{LC}} \bar{\boldsymbol{\alpha}}_1^\top \bar{\boldsymbol{\varphi}}_{1i}}{\sqrt{N_{\mathcal{LC}}}} - \boldsymbol{h}^\top \boldsymbol{V}_{\bar{\boldsymbol{\gamma}}}^{-1} \frac{\sum_{t_i=C} \Phi_i}{\sqrt{N_{\mathcal{LC}}}} \\
&+ \frac{\sqrt{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC},\mathcal{UC}} \bar{\boldsymbol{\alpha}}_2^\top \bar{\boldsymbol{\varphi}}_{2i}}{N_{\mathcal{LC}}+N_{\mathcal{UC}}} - \frac{\sum_{t_i=\mathcal{LC}} \bar{\boldsymbol{\alpha}}_2^\top \bar{\boldsymbol{\varphi}}_{2i}}{\sqrt{N_{\mathcal{LC}}}} + o_P(1)
\end{aligned}$$

Define $\mathrm{R}(\widetilde{\boldsymbol{\varphi}}_1) = \boldsymbol{h}^\top \widetilde{\boldsymbol{S}} - \frac{N_{IC}}{N_{IC}+N_{\mathcal{LC}}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_1$. We also have $\sqrt{N_{\mathcal{LC}}}(\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_2 - \boldsymbol{h}^\top \bar{\boldsymbol{\gamma}}) \to N(0, \Sigma_{\widetilde{\boldsymbol{\gamma}}_2})$ where $\Sigma_{\widetilde{\boldsymbol{\gamma}}_2} = \frac{\mathbb{E}\left(\mathrm{R}(\widetilde{\boldsymbol{\varphi}}_1) - \tilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_2\right)^2}{N_{\mathcal{LC}}} + \frac{\mathbb{E}\left(\tilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_2\right)^2}{N_{\mathcal{LC}}+N_{\mathcal{UC}}} + \frac{N_{IC}\mathbb{E}\left(\tilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_1\right)^2}{(N_{\mathcal{LC}}+N_{IC})^2}$. We plug the expression of $\widetilde{\boldsymbol{\gamma}}_1$ into the expression of $\widetilde{\boldsymbol{\varphi}}_2$. Then we have

$$\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_2 = \boldsymbol{h}^\top \bar{\boldsymbol{\gamma}} + \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \mathrm{R}(\widetilde{\boldsymbol{\varphi}}_{1i}) + \frac{1}{N_{\mathcal{LC}}+N_{\mathcal{LM}}} \sum_{t_i=IC} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i} - \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i} + \frac{1}{N_{\mathcal{LC}}+N_{\mathcal{UC}}} \sum_{t_i=\mathcal{LC},\mathcal{UC}} \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i}$$

And

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_1) &= \mathbb{E}[(\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \mathrm{R}(\widetilde{\boldsymbol{\varphi}}_{1i}) + \frac{1}{N_{\mathcal{LC}}+N_{\mathcal{LM}}} \sum_{t_i=IC} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i} - \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i} + \frac{1}{N_{\mathcal{LC}}+N_{\mathcal{UC}}} \sum_{t_i=\mathcal{LC},\mathcal{UC}} \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i})^2] \\
&= \mathbb{E}\{[\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} (\mathrm{R}(\widetilde{\boldsymbol{\varphi}}_{1i}) - \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i}) + \frac{1}{N_{\mathcal{LC}}+N_{\mathcal{LM}}} \sum_{t_i=IC} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i} + \frac{1}{N_{\mathcal{LC}}+N_{\mathcal{UC}}} \sum_{t_i=\mathcal{LC},\mathcal{UC}} \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i}]^2\}
\end{aligned}$$

Remember that $\widetilde{\boldsymbol{\alpha}}_2$ is the solution to linear regression between $\mathrm{R}(\widetilde{\boldsymbol{\varphi}}_{1i})$ and $\widetilde{\boldsymbol{\varphi}}_{2i}$. Thus, $\mathrm{R}(\widetilde{\boldsymbol{\varphi}}_{1i}) - \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i}$ should be orthogonal to $\widetilde{\boldsymbol{\varphi}}_{2i}$. All the individuals are independent. We have

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_2) &= \mathbb{E}\{[\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} (\mathrm{R}(\widetilde{\boldsymbol{\varphi}}_{1i}) - \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i}) + \frac{1}{N_{\mathcal{LC}}+N_{\mathcal{LM}}} \sum_{t_i=IC} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i} + \frac{1}{N_{\mathcal{LC}}+N_{\mathcal{UC}}} \sum_{t_i=\mathcal{LC},\mathcal{UC}} \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i}]^2\} \\
&= \frac{\mathbb{E}(\mathrm{R}(\widetilde{\boldsymbol{\varphi}}_1) - \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_2)^2}{N_{\mathcal{LC}}} + \frac{\mathbb{E}(\widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_2)^2}{N_{\mathcal{LC}}+N_{\mathcal{UC}}} + \frac{N_{\mathcal{LM}}\mathbb{E}(\widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_1)^2}{(N_{\mathcal{LC}}+N_{\mathcal{LM}})^2}
\end{aligned}$$

Similarly, the variance of $\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}$ can be estimated

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{h}^\top \widetilde{\boldsymbol{\gamma}}_1) &= \mathbb{E}\{[\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} (\mathrm{R}(\widetilde{\boldsymbol{\varphi}}_{1i}) - \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i}) + \frac{1}{N_{\mathcal{LC}}+N_{\mathcal{LM}}} \sum_{t_i=IC} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i} + \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_{2i}]^2\} \\
&= \frac{\mathbb{E}(\mathrm{R}(\widetilde{\boldsymbol{\varphi}}_1) - \widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_2)^2}{N_{\mathcal{LC}}} + \frac{\mathbb{E}(\widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_2)^2}{N_{\mathcal{LC}}} + \frac{N_{\mathcal{LM}}\mathbb{E}(\widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_1)^2}{(N_{\mathcal{LC}}+N_{\mathcal{LM}})^2}
\end{aligned}$$

We can also see our proposed estimator has guaranteed efficiency since $\frac{\mathbb{E}\left(\widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_2\right)^2}{N_{\mathcal{LC}}} > \frac{\mathbb{E}\left(\widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_2\right)^2}{N_{\mathcal{LC}}+N_{\mathcal{UC}}}$. And the variance ratio would depend on $\mathbb{E}\left(\widetilde{\boldsymbol{\alpha}}_2^\top \widetilde{\boldsymbol{\varphi}}_2\right)^2$ and sample size $N_{\mathcal{LC}}, N_{\mathcal{UC}}$.

# E Lemma

**Lemma 3.** *If* $\boldsymbol{X}_n \xrightarrow{p} \boldsymbol{X}$ *and* $\boldsymbol{Y}_n \xrightarrow{p} \boldsymbol{Y}$*, then* $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow{p} \boldsymbol{X} + \boldsymbol{Y}$

Proof:

$$A_n = \left\{|X_n - X| < \frac{\varepsilon}{2}\right\} \quad B_n = |Y_n - Y\ | < \frac{\varepsilon}{2}\Big\}$$
$$P\left(A_n \cap B_n\right) \geq P\left(A_n\right) + P\left(B_n\right) - 1$$
$$C_n = \{|X_n - X| + |Y_n - Y| < \varepsilon\}$$
$$D_n = |\ X_n - X + Y_n - Y\ | < \varepsilon$$
$$A_n \cap B_n \subset C_n \subset D_n \quad \lim P\left(D_n\right) \geq \lim P\left(A_n \cap B_n\right) = 1$$

**Lemma 4.** *If* $\boldsymbol{X}_n \xrightarrow{p} \boldsymbol{X}$ *and* $\boldsymbol{Y}_n \xrightarrow{p} \boldsymbol{Y}$*, then* $\boldsymbol{X}_n \times \boldsymbol{Y}_n \xrightarrow{p} \boldsymbol{X} \times \boldsymbol{Y}$

Proof:

$$|X_n - X|\ |\ Y_n - Y\ | < \frac{\varepsilon^2}{3} < \frac{\varepsilon}{3} \text{ when } \varepsilon \to 0$$
$$|Y|\,|X_n - X| < \frac{\varepsilon}{3}$$
$$|X|\,|Y_n - Y| < \frac{\varepsilon}{3}$$

**Lemma 5.** $\widetilde{\boldsymbol{\alpha}}_1$ *is asymptotically normality to* $\bar{\boldsymbol{\alpha}}_1$ *under the* $\bar{\boldsymbol{\varphi}}_1$ *condition, i.e.,* $\boldsymbol{\varphi}_1$ *is true, and*

$$(\widetilde{\boldsymbol{\alpha}}_1 - \bar{\boldsymbol{\alpha}}_1) = \bar{\boldsymbol{V}}^{-1} \frac{1}{N_{\mathcal{LC}}} \sum_{t_i = C} \bar{\Psi}_i + o_P(\frac{1}{\sqrt{N_{\mathcal{LC}}}})$$

*where*

$$\bar{\Psi}_i = \bar{\boldsymbol{\varphi}}_{1i}\{\boldsymbol{h}^\top \bar{\boldsymbol{S}}_i - \bar{\boldsymbol{\alpha}}_1^\top \bar{\boldsymbol{\varphi}}_{1i}\}, \quad \bar{\boldsymbol{V}} = \frac{1}{N_{\mathcal{LC}}} \sum_{t_i = C} \bar{\boldsymbol{\varphi}}_{1i}^{\otimes 2}$$

**Lemma 6.** $\widetilde{\boldsymbol{\alpha}}_2$ *is asymptotically normality to* $\bar{\boldsymbol{\alpha}}_2$ *under the* $\bar{\boldsymbol{\varphi}}_2$ *condition, i.e.,* $\boldsymbol{\varphi}_2$ *is true, and*

$$(\widetilde{\boldsymbol{\alpha}}_2 - \bar{\boldsymbol{\alpha}}_2) = \bar{\boldsymbol{\mathcal{V}}}^{-1} \frac{1}{N_{\mathcal{LC}}} \sum_{t_i = C} \bar{\rho}_i + o_P(\frac{1}{\sqrt{N_{\mathcal{LC}}}})$$

*where*

$$\bar{\rho}_i = \bar{\boldsymbol{\varphi}}_{2i}\{\boldsymbol{h}^\top \bar{\boldsymbol{S}}_i - \frac{N_{\mathcal{LM}}}{N_{\mathcal{LM}} + N_{\mathcal{LC}}} \bar{\boldsymbol{\alpha}}_1^\top \bar{\boldsymbol{\varphi}}_{1i} - \bar{\boldsymbol{\alpha}}_2^\top \bar{\boldsymbol{\varphi}}_{2i}\}, \quad \bar{\boldsymbol{\mathcal{V}}} = \frac{1}{N_{\mathcal{LC}}} \sum_{t_i = C} \bar{\boldsymbol{\varphi}}_{2i}^{\otimes 2}$$

**Proof of Lemma 1** $\widetilde{\boldsymbol{\varphi}}_1$ is the integration and has the form $\widetilde{\boldsymbol{\varphi}}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y) = \int \widetilde{\boldsymbol{S}} \widetilde{P}(\boldsymbol{X}_{\mathcal{P}_m} \mid \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y)\, d\boldsymbol{X}_{\mathcal{P}_m}$, where $\widetilde{\boldsymbol{S}}$ and $\widetilde{P}(\boldsymbol{X}_{\mathcal{P}_m} \mid \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y)\, d\boldsymbol{X}_{\mathcal{P}_m}$ are functions respective to parameter $\widetilde{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\eta}}$. Thus, $\widetilde{\boldsymbol{\varphi}}_1$ can be re-written in the form $\widetilde{\boldsymbol{\varphi}}_1 = \boldsymbol{I}(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\eta}})$ where $\boldsymbol{I}(\cdot)$ represents integration. We consider the Taylor expansion of $\boldsymbol{I}$:

$$\widetilde{\boldsymbol{\varphi}}_1 = \boldsymbol{I}(\bar{\boldsymbol{\gamma}}, \bar{\boldsymbol{\eta}}) + \dot{\boldsymbol{I}}_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\gamma}}, \bar{\boldsymbol{\eta}})(\widetilde{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}) + \dot{\boldsymbol{I}}_{\boldsymbol{\eta}}(\bar{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\eta}})(\widetilde{\boldsymbol{\eta}} - \bar{\boldsymbol{\eta}}) + \ddot{\boldsymbol{I}}_{\boldsymbol{\gamma}\boldsymbol{\eta}}(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\eta}})(\widetilde{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}})(\widetilde{\boldsymbol{\eta}} - \bar{\boldsymbol{\eta}})$$

The Integration is divided into four parts. The first part is constant. The second and third parts are correlated to the consistency of two parameters respectively. The fourth part is the product of two consistent parts. By **lemma4** we conclude the fourth part is still consistent. We might conclude that the error rate of the sequences $(\boldsymbol{I}(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\eta}}) - \boldsymbol{I}(\bar{\boldsymbol{\gamma}}, \bar{\boldsymbol{\eta}}))$ is $o_P(1)$ since this sequences is the summation of three sequences whose error rates are bounded $o_P(N_{\mathcal{LC}}^{-\frac{1}{2}})$ and $o_P(1)$. Recall $\boldsymbol{I}(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\eta}})$ is the integration of $\widetilde{\boldsymbol{S}}$ and $\widetilde{P}(\boldsymbol{X}_{\mathcal{P}_m} \mid \boldsymbol{X}_{\mathcal{P}_{m^c}}, Y = 0, 1)$, we should also assume the continuity of two functions to ensure the existence of derivative for the integration, i.e., $\widetilde{\boldsymbol{S}}, \widetilde{P} \in C^0$. We should also note that we apply the Monte-Carlo method to approximate the integration

numerically with the belief that when the resample size is large enough, the error between our numerical estimation ($\boldsymbol{I}_{mc}$) and true integration can be $o_P(1)$. $\boldsymbol{I}_{mc} - \boldsymbol{I}(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\eta}})$ is $o_P(1)$, and $\boldsymbol{I}(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\eta}}) - \boldsymbol{I}(\bar{\boldsymbol{\gamma}}, \bar{\boldsymbol{\eta}})$ is $o_P(1)$. Thus, $\boldsymbol{I}_{mc} - \boldsymbol{I}(\bar{\boldsymbol{\gamma}}, \bar{\boldsymbol{\eta}})$ is $o_P(1)$.

**Proof of Lemma 2** We begin by showing that $\widetilde{\boldsymbol{\alpha}}_1 \xrightarrow{p} \bar{\boldsymbol{\alpha}}_1$. We note that $\widetilde{\boldsymbol{\alpha}}_1$ is the solution to

$$\widetilde{\mathbf{U}}_{N_{\mathcal{LC}}}(\boldsymbol{\alpha}_1) = \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\varphi}}_1(\boldsymbol{X}_{i\mathcal{M}^c}, Y_i)\{\boldsymbol{h}^\top \widetilde{\boldsymbol{S}}_i - \boldsymbol{\alpha}_1^\top \widetilde{\boldsymbol{\varphi}}_1(\boldsymbol{X}_{i\mathcal{M}^c}, Y_i)\} = \mathbf{0}$$

It follows by the ULLN since all $\widetilde{\boldsymbol{S}}$ and $\widetilde{\boldsymbol{\varphi}}_1$ are continuously differentiable in $\widetilde{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\eta}}$ and uniformly bounded. The consistency of $\widetilde{\boldsymbol{\alpha}}_1$ for $\bar{\boldsymbol{\alpha}}_1$ them follows from the fact that $\widetilde{\boldsymbol{\gamma}}$ converge in probability to $\bar{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\eta}}$ converge in probability to $\bar{\boldsymbol{\eta}}$. To establish the asymptotic distribution of $\widetilde{\boldsymbol{\alpha}}_1$, we apply Theorem **5.21 of Van der Vaart [2000]** to obtain the Taylor expansion:

$$(\widetilde{\boldsymbol{\alpha}}_1 - \bar{\boldsymbol{\alpha}}_1) = \bar{\boldsymbol{V}}^{-1} \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=C} \bar{\Psi}_i + o_P(\frac{1}{\sqrt{N_{\mathcal{LC}}}})$$

where

$$\bar{\Psi}_i = \bar{\boldsymbol{\varphi}}_{1i}\{\boldsymbol{h}^\top \bar{\boldsymbol{S}}_i - \bar{\boldsymbol{\alpha}}_1^\top \bar{\boldsymbol{\varphi}}_{1i}\}, \quad \bar{\boldsymbol{V}} = \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=C} \bar{\boldsymbol{\varphi}}_{1i}^{\otimes 2}$$

It then follows the classical central limit theorem that $\sqrt{N_{\mathcal{LC}}}(\widetilde{\boldsymbol{\alpha}}_1 - \bar{\boldsymbol{\alpha}}_1) \to N(0, \Sigma_{\widetilde{\boldsymbol{\alpha}}_1})$ where $\Sigma_{\widetilde{\boldsymbol{\alpha}}_1} = \mathbb{E}\bar{\boldsymbol{\varphi}}_{1i}^{\otimes 2}$

**Proof of Lemma 3** $\widetilde{\boldsymbol{\varphi}}_2$ is the integration and has the simplified form

$$\widetilde{\boldsymbol{\varphi}}_2(\boldsymbol{X}) = \widetilde{P}(Y=1|\boldsymbol{X})\widetilde{\boldsymbol{\varphi}}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y=1) + \widetilde{P}(Y=0|\boldsymbol{X})\widetilde{\boldsymbol{\varphi}}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y=0)$$

where $\widetilde{\boldsymbol{\varphi}}_1$ is the function respective to parameter $\widetilde{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\eta}}$. $\widetilde{P}(Y|\boldsymbol{X})$ is the function respective to parameter $\widetilde{\boldsymbol{\gamma}}$. The Integration is divided into two parts and each part is the product of proved consistent items. By **lemma2** we conclude $\widetilde{\boldsymbol{\varphi}}_2$ is consistent and is $o_P(1)$ since this sequence is the summation of two sequences whose error rates are bounded $o_P(1)$. We should also note that we apply the Monte-Carlo method to approximate the integration numerically with the belief that when the resample size is large enough, the error between our numerical estimation ($\boldsymbol{I}_{mc}$) and true integration can be $o_P(1)$. $\boldsymbol{I}_{mc} - \widetilde{\boldsymbol{\varphi}}_2$ is $o_P(1)$, and $\widetilde{\boldsymbol{\varphi}}_2 - \bar{\boldsymbol{\varphi}}_2$ is $o_P(1)$. Thus, $\boldsymbol{I}_{mc} - \bar{\boldsymbol{\varphi}}_2$ is $o_P(1)$.

**Proof of Lemma 4** We begin by showing that $\widetilde{\boldsymbol{\alpha}}_2 \xrightarrow{p} \bar{\boldsymbol{\alpha}}_2$. We note that $\widetilde{\boldsymbol{\alpha}}_2$ is the solution to

$$\mathbf{U}_{N_{\mathcal{LC}}}(\boldsymbol{\alpha}_2) = \mathbb{E}\boldsymbol{\varphi}_2(\boldsymbol{X})\{\boldsymbol{h}^\top \boldsymbol{S} - \frac{N_{\mathcal{LM}}}{N_{\mathcal{LM}} + N_{\mathcal{LC}}} \boldsymbol{\alpha}_1^\top \boldsymbol{\varphi}_1(\boldsymbol{X}_{\mathcal{P}_{m^c}}, Y) - \boldsymbol{\alpha}_2^\top \boldsymbol{\varphi}_2(\boldsymbol{X})\} = \mathbf{0}$$

The $\boldsymbol{h}^\top \boldsymbol{S} - \frac{N_{\mathcal{LM}}}{N_{\mathcal{LM}}+N_{\mathcal{LC}}} \boldsymbol{\alpha}_1^\top \boldsymbol{\varphi}_1$ comes from the estimation in $\widetilde{\boldsymbol{\gamma}}_1$. Focus on the complete data set

$$\begin{aligned}
&\frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \boldsymbol{h}^\top \widetilde{\boldsymbol{S}}_i - \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i} + \frac{1}{N_{\mathcal{LC}} + N_{\mathcal{LM}}} \sum_{t_i=\mathcal{LC}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i} \\
&= \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} [\boldsymbol{h}^\top \widetilde{\boldsymbol{S}}_i + (\frac{N_{\mathcal{LC}}}{N_{\mathcal{LC}} + N_{\mathcal{LM}}} - 1)\widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}] \\
&= \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=\mathcal{LC}} [\boldsymbol{h}^\top \widetilde{\boldsymbol{S}}_i - \frac{N_{\mathcal{LM}}}{N_{\mathcal{LC}} + N_{\mathcal{LM}}} \widetilde{\boldsymbol{\alpha}}_1^\top \widetilde{\boldsymbol{\varphi}}_{1i}]
\end{aligned}$$

$\widetilde{\boldsymbol{\alpha}}_2$ follows by the ULLN since all $\widetilde{\boldsymbol{S}}$ and $\widetilde{\boldsymbol{\varphi}}_1$ are continuously differentiable in $\widetilde{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\eta}}$ and uniformly bounded. The consistency of $\widetilde{\boldsymbol{\alpha}}_2$ for $\bar{\boldsymbol{\alpha}}_2$ them follows from the fact that $\widetilde{\boldsymbol{\gamma}}$ converge in probability to $\bar{\boldsymbol{\gamma}}$, $\widetilde{\boldsymbol{\alpha}}_1$ converge in probability to $\bar{\boldsymbol{\alpha}}_1$ and $\widetilde{\boldsymbol{\eta}}$ converge in probability to $\bar{\boldsymbol{\eta}}$. To establish the asymptotic distribution of $\widetilde{\boldsymbol{\alpha}}_1$, we apply Theorem **5.21 of Van der Vaart [2000]** to obtain the Taylor expansion:

$$(\widetilde{\boldsymbol{\alpha}}_2 - \bar{\boldsymbol{\alpha}}_2) = \bar{\boldsymbol{\mathcal{V}}}^{-1} \frac{1}{N_{\mathcal{LC}}} \sum_{t_i=C} \bar{\rho}_i + o_P(\frac{1}{\sqrt{N_{\mathcal{LC}}}})$$

where

$$\bar{\rho}_i = \bar{\boldsymbol{\varphi}}_{2i}\{\boldsymbol{h}^\top \bar{\boldsymbol{S}}_i - \frac{N_{\mathcal{LM}}}{N_{\mathcal{LM}} + N_{\mathcal{LC}}}\bar{\boldsymbol{\alpha}}_1^\top \bar{\boldsymbol{\varphi}}_{1i} - \bar{\boldsymbol{\alpha}}_2^\top \bar{\boldsymbol{\varphi}}_{2i}\}, \quad \bar{\boldsymbol{\nu}} = \frac{1}{N_{\mathcal{LC}}}\sum_{t_i=C} \bar{\boldsymbol{\varphi}}_{2i}^{\otimes 2}$$

It then follows the classical central limit theorem that $\sqrt{N_{\mathcal{LC}}}(\widetilde{\boldsymbol{\alpha}}_2 - \bar{\boldsymbol{\alpha}}_2) \to N(0, \Sigma_{\widetilde{\boldsymbol{\alpha}}_2})$ where $\Sigma_{\widetilde{\boldsymbol{\alpha}}_2} = \mathbb{E}\bar{\boldsymbol{\varphi}}_{2i}^{\otimes 2}$

# F Misspecified Model

$$logit(P(Y = 1|\boldsymbol{X})) = \boldsymbol{X}\boldsymbol{\gamma} + X_1^2 + X_2^2$$

We also want to show our proposed methods efficient even in misspecifed model settings. We got the $P(Y_i = 1|\boldsymbol{X_i})$ from the misspecified model above for each observation, then we generated $Y_i$ from $B(P(Y_i = 1|\boldsymbol{X_i}))$.

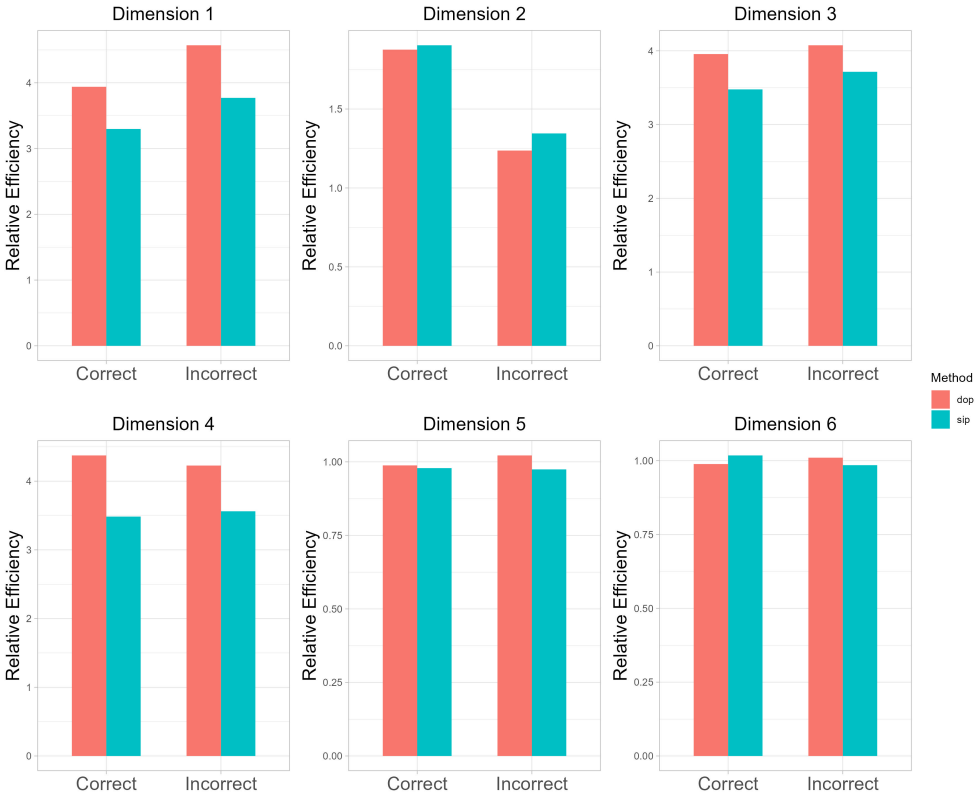


Figure 6: Relative Efficiency of proposed estimator from misspecified model

We could see from Figure 4, there is no much difference between the relative efficiency of Sip and Dop from misspecified model compared with the correctly specified model.

# G AUC

AUC for lab 7
    AUC for score 8
    lab name reference 9

Table 7: Estimated parameters

| AUC | without lab | with lab |
|---|---|---|
| hormone in psychiatric | 0.66 | 0.675 |
| microcytes in lung diseases | 0.72 | 0.739 |

Table 8: AUC for score

| AUC1 | AUC2 | phenotype | target lab | similar lab |
|---|---|---|---|---|
| 0.701 | 0.746 | heart disease | 50993 | 51279,51274 |
| 0.783 | 0.845 | obesity | 50801 | 51279,50802,50816,50820,50821,151274 |
| 0.640 | 0.837 | lung disease | 50801 | 51279,50802,50816,50820,50821,51274 |
| 0.673 | 0.706 | lung disease | 51260 | 51279,51137,51233,51246,51252,51267 |
| 0.759 | 0.822 | alcohol abuse | 50801 | 51279,50802,50816,50820,50821,51274 |

Table 9: lab referennce

| itemid | label | fluid | category | loinc code |
|---|---|---|---|---|
| 50801 | Alveolar-arterial Gradient | Blood | Blood Gas | $19991-9$ |
| 50802 | Base Excess | Blood | Blood Gas | $11555-0$ |
| 50816 | Oxygen | Blood | Blood Gas | $19994-3$ |
| 50820 | pH | Blood | Blood Gas | $11558-4$ |
| 50821 | pO2 | Blood | Blood Gas | $11556-8$ |
| 50993 | Thyroid Stimulating Hormone | Blood | Chemistry | $3016-3$ |
| 51137 | Anisocytosis | Blood | Hematology | $702-1$ |
| 51233 | Hypochromia | Blood | Hematology | $728-6$ |
| 51246 | Macrocytes | Blood | Hematology | $738-5$ |
| 51252 | Microcytes | Blood | Hematology | $741-9$ |
| 51260 | Ovalocytes | Blood | Hematology | $774-0$ |
| 51267 | Poikilocytosis | Blood | Hematology | $779-9$ |
| 51274 | PT | Blood | Hematology | $5902-2$ |
| 51279 | Red Blood Cells | Blood | Hematology | $789-8$ |