**Title**
Targeted Learning for Capture Recapture Models and Treatment Effect Estimation



**Author**
You, Yue

**Publication Date**
2021

Peer reviewed|Thesis/dissertation

# Targeted Learning for Capture Recapture Models and Treatment Effect Estimation

by

Yue You

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Alan Hubbard, Co-chair
Professor Mark van der Laan, Co-chair
Professor Justin Remais

Spring 2021

Targeted Learning for Capture Recapture Models and Treatment Effect Estimation

Abstract

Targeted Learning for Capture Recapture Models and Treatment Effect Estimation

by

Yue You

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Alan Hubbard, Co-chair

Professor Mark van der Laan, Co-chair

This dissertation develops modern statistical methods, targeted maximum likelihood estimation (TMLE) and ensemble machine learning, for two common problems in epidemiology and public health: 1) estimating the population size based on capture recapture designs, and 2) estimating the continuous and discrete treatment effect of multiple exposures. For the first problem, We proposed novel target parameters for each identification assumption and robust estimators based on TMLE, provided the efficient influence curves for each the parameter, proved the statistical properties of the estimators, and applied them to data collected from national-level infectious disease surveillance systems. We also used simulations with identification assumption violations to test the reliability of the estimation. For the second problem, we developed our target parameter and TMLE estimators and applied them to various types of empirical data collected by community-level follow-up surveys and state-level electronic health record data. We provided simulations and sensitivity analysis to show the performance of the TMLE estimators compared to existing ones. In chapter 1, we gave an introduction to TMLE, the road map of targeted learning, and a more detailed summary of the following chapters. In chapter 2, we developed a novel approach to estimate the population size based on capture recapture designs and evaluated the estimation reliability. In chapter 3, we utilized the targeted learning approach to assess the performance of a diabetes care program on glycemic control of type 2 diabetes patients, and identified patient subgroups with most successful treatment effects. In chapter 4, we proposed a robust variable importance measure based on TMLE and applied it in estimating the transmission effects of mother's eating behavior on the next generation.

To Qinqi, mom and dad.

# Contents

# List of Figures

# List of Tables

# Acknowledgments

During my years at Berkeley, I am very fortunate to be advised by professor Alan Hubbard and professor Mark van der Laan. Throughout my coursework, my doctoral research and during the writing of my dissertation, Alan and Mark helped me exploring my academic interests, digging into the challenging problems and patiently trained me how to read, write, and think as a scholar. I could not express how much I appreciate their insightful suggestions and full support during the last five years. 2020 is a very hard year for us all, and I want to give special thanks to Alan for going extra mile to care for my condition during the pandemic. I am also extremely grateful to professor Justin Remais, who welcomed me to his research group, guided me to many aspects of the surveillance data project, which inspired the major part of my dissertation. He always gives me wise advice and comments, and helped me so much in organizing and writing up the dissertation.

I am also very grateful to the community of the School of Public Health, Statistics Department and UCSF for all the research opportunities I had. I have learnt so much from the faculty and researchers and will always treasure my amazing experience working with professor Maya Petersen, professor Barbara Laraia, professor Nick Jewell, professor Katrina Abuabara, and professor Haiyan Huang. I want to thank Svetlana Doubova and professor Stefano Bertozzi for the exciting experience to work with IMSS and visit Mexico. And I really appreciate the GSI opportunity with Maureen Lahiff. The happy time with my fellow students and friends is also a precious experience. I will always cherish the memory with Qu Cheng, Rachael Phillips, Nima Hejazi, Wilson Cai, Jonathan Levy, Philip Collender, Waverly Wei, Chi Zhang, Cheng Ju and Ivana Malenica for the wonderful time we spent at Berkeley.

Part of the materials presented here have been published elsewhere. Materials in chapter 2 is co-authored with Mark van der Laan, Philip Collender, Qu Cheng, Zhiyue Hu, Alan Hubbard, Nick Jewell, Robin Mejia, and Justin Remais, materials in chapter 3 appeared on BMC Medical Informatics and Decision Making as "Application of machine learning methodology to assess the performance of DIABETIMSS program for patients with type 2 diabetes in family medicine clinics in Mexico", co-authored with Svetlana V. Doubova , Diana Pinto-Masis, Ricardo Perez-Cuevas, Victor Hugo Borja-Aburto and Alan Hubbard, and materials in chapter 4 is co-authored with Alan Hubbard and Babara Laraia. I sincerely thank each of co-authors for their contributions and permission to include the work in my dissertation. This dissertation was partly supported by NIH grant R01AI125842 and NIH grant 2R01AI074345.

# Chapter 1

# Introduction

## 1.1 Background

In this section we briefly reviewed the elements of targeted maximum likelihood estimation (TMLE) for general estimation problem with an example of estimating the average treatment effect (ATE) [1] in observational studies. For more details we refer to [2].

### Introduction to targeted maximum likelihood estimation (TMLE)

The idea of targeted maximum likelihood estimator (TMLE) was introduced by van der Laan and Rubin [3]. TMLE is a doubly robust, maximum-likelihood–based estimation method that includes a secondary "targeting" step to optimize the bias-variance trade-off for the target parameter. A detailed step-by-step guide to TMLE can be found in Chapter 4 of [2], and a simple motivating example can be found in [4]. One common use case for TMLE is to estimate the ATE in observational studies. Let us denote the data structure of such an observational study as $O = (W, A, Y) \sim P_0$, where $W$ represents the covariates, $A$ represents a binary exposure or treatment, $Y$ represents the outcome, $P_0$ represents the true probability distribution of $O$, and the nonparametric or semi-parametric statistical model $\mathcal{M}$ represents the set of possible probability distributions for $P_0$. The additive causal effect target parameter $\Psi(P_0) = E_{W,0}[E_0(Y|A = 1, W) - E_0(Y|A = 0, W)]$. There are four steps to implement a targeted maximum likelihood estimator for $\Psi$:

1. Estimate outcome mechanism (initial estimate of $E(Y|A, W)$): we could use a variety of machine learning algorithms in this step, such as linear regressions, LASSO [5] and random forest [6]. Because we do not know the true distribution of the empirical data, selecting an optimal algorithm can be challenging. We recommend using an ensemble machine learning method which allows researchers to include a collection of algorithms and report the optimal one or a combination of algorithms based on pre-specified criteria. We often use an ensemble learner called Super Learner [7] in this step. Super learner is a data-adaptive ensemble learner which combines user-input

algorithms through weighting to minimize the cross-validated mean squared error (or other risk measure) [4]. We denote the estimated value of $E(Y|A=1,W)$ as $\hat{Y}_1$, and the estimated value of $E(Y|A=0,W)$ as $\hat{Y}_0$.

2. estimate exposure(treatment) mechanism $P(A=1|W)$: for this step, we could also use the algorithms discussed in the first step (such as Super Learner).

3. update the initial estimate of $E(Y|A,W)$: for ATE, a "clever covariate" is defined as $H_a(A=a,W) = \frac{\mathbb{1}(A=1)}{P_n(A=1|W)} - \frac{\mathbb{1}(A=0)}{P_n(A=0|W)}$. The clever covariate is derived from the efficient influence function, and is used to define a parametric working sub-model that fluctuates the initial estimator to reduce bias for the target parameter. We fit a logistic regression of the outcome $Y$ on $H_a$ using as a fixed intercept the offset $logit(\hat{Y}_a) : logit(E^*(Y|A,W)) = logit(\hat{Y}_a) + \epsilon H_a(A,W)$, and compute $\hat{\epsilon}$ which minimizes the empirical loss function. Then we generate the updated estimates as $logit(\hat{Y}_a^*) = logit(\hat{Y}_a) + \hat{\epsilon} H_a$.

4. Generate targeted estimate of target parameter: the targeted maximum likelihood estimator of the target parameter (which is ATE in this case) is $\psi_n^{TMLE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_1^* - \hat{Y}_0^*)$.

In all the following three chapters, we developed the TMLE for their corresponding target parameters. The target parameter in chapter 3 is the ATE, and the TMLE in chapter 3 is constructed in a similar way to the example above. In chapter 4, the target parameter involves the ATE of a continuous treatment variable, and the updating step of the TMLE utilizes a different efficient influence function in step 3. In chapter 1, the target parameter is not related to ATE, and the implementation of the TMLE only has three steps: 1) generate an initial estimate, 2) construct a least favorable parametric model through the initial estimator and update the initial estimator through an iterative process, and 3) generate targeted estimate of target parameter. The implementation specifications of the three TMLE's are reported in the corresponding chapters.

All the three targeted maximum likelihood estimators developed in the dissertation share the following common advantages. First, TMLE has great flexibility to incorporate a variety of statistical learning algorithms as its initial estimator. By incorporating various algorithms, the TMLE is less vulnerable to the bias induced by model mis-specifications, which is a practical concerns for complex observational data. Second, TMLE is a plug-in or substitution estimator, which are know to be more robust to outliers and sparsity than are non-substitution estimators. Third, TMLE is an asymptotically efficient estimator under certain assumptions. When ATE is the target parameter, TMLE is an asymptotically efficient when both the outcome and exposure mechanisms are consistently estimated. Finally, with ATE as the target parameter, TMLE is a doubly robust and will yield unbiased estimates if either $E(Y|A,W)$ or $P(A=1|W)$ is consistently estimated (e.g., correctly specified in the case of parametric regression). If the outcome regression is not consistently estimated,

the final ATE estimate will still be unbiased if the exposure mechanism is consistently estimated. Conversely, if the outcome is consistently estimated, the targeting step will preserve this unbiasedness and may remove finite sample bias [2, 4].

## 1.2 Road map and chapter summaries

In all the following chapters we followed the road map proposed by Rose and van der Laan [2]. The road map has three components and the language is shared by all three chapters. First, we define the research question. The data are n identically and independently distributed observations of random variable $O$, which has probability distribution $P_0$. The statistical model $\mathcal{M}$ is a set of possible distributions of $O$. $P_0 \in \mathcal{M}$. The target parameter or parameter of interest $\Psi(P_0)$ is a particular feature of $P_0$ where $\Psi$ maps the probability distribution $P_0$ into the target parameter. Second, We estimate the target parameter. We build an initial estimator of the relevant part $Q_0$ of $P_0$ using machine learning algorithm Super Learner. Then we update the initial fit in a step targeted toward making an optimal bias-variance tradeoff for the target parameter, now denoted as $\Psi(Q_0)$, instead of the overall probability distribution. Last, we provide inference and interpretation for our estimation. In this step, standard errors are calculated for the estimator of the target parameter using the influence curve or resampling-based methods to assess the uncertainty in the estimator. The target parameter can be interpreted as a purely statistical parameter or as a causal parameter under possible additional nontestable assumptions in our model [2].

In chapter 2, we developed a novel approach to estimate the population size based on capture recapture designs and evaluated the estimation reliability. In particular, we proposed a modern method to estimate population size based on capture-recapture designs of K samples. The observed data is formulated as a sample of n i.i.d. K-dimensional vectors of binary indicators – where the k-th component of each vector indicates the subject being caught by the k-th sample – such that only subjects with nonzero capture vectors are observed. The target quantity is the unconditional probability of the vector being nonzero across both observed and unobserved subjects. We covered models assuming a single general constraint on the K-dimensional distribution such that the target quantity is identified and the statistical model is unrestricted. We presented solutions for general linear constraints, as well as constraints commonly assumed to identify capture-recapture models, including no K-way interaction in linear and log-linear models, independence or conditional independence. We demonstrated that the choice of constraint(identification assumption) has a dramatic impact on the value of the estimand, showing that it is crucial that the constraint is known to hold by design. For the commonly assumed constraint of no K-way interaction in a log-linear model, the statistical target parameter is only defined when each of the $2^K - 1$ observable capture patterns is present, and therefore suffers from the curse of dimensionality. We proposed a targeted MLE based on undersmoothed lasso model to smooth across the cells while targeting the fit towards the single valued target parameter of interest. For each identification assumption, we provided simulated inference and confidence intervals to

assess the performance on the estimator under correct and incorrect identifying assumptions. We applied the proposed method, alongside existing estimators, to estimate prevalence of a parasitic infection using multi-source surveillance data from a region in southwestern China, under the four identification assumptions.

In chapter 3, we utilized a machine-learning-based targeted learning approach to assess the performance of a diabetes care program on glycemic control of type 2 diabetes patients, and identified patient subgroups with most successful treatment effects. Specifically, we analyzed the EHR and laboratory databases from the year 2012 to 2016 of T2D patients from six family medicine clinics (FMCs) delivering the DIABETIMSS program, and five FMCs providing routine care. The primary outcome was glycemic control. The study covariates included: patient sex, age, anthropometric data, history of glycemic control, diabetic complications and comorbidity. We measured the effects of DIABETIMSS program through 1) simple unadjusted mean differences; 2) adjusted via standard logistic regression and 3) adjusted via targeted machine learning. We treated the data as a serial cross-sectional study, conducted a standard principal components analysis to explore the distribution of covariates among clinics, and performed regression tree on data transformed to use the prediction model to identify patient sub-groups in whom the program was most successful. To explore the robustness of the machine learning approaches, we conducted a set of simulations and the sensitivity analysis with process-of-care indicators as possible confounders. The results showed that the impact of DIABETIMSS ranged, among clinics, from 2 to 8% improvement in glycemic control, with an overall (pooled) estimate of 5% improvement. T2D patients with fewer complications have more significant benefit from DIABETIMSS than those with more complications. At the FMC's delivering the conventional model the predicted impacts were like what was observed empirically in the DIABETIMSS clinics. The sensitivity analysis did not change the overall estimate average across clinics. In conclusion, the DIABETIMSS program had a small, but significant increase in glycemic control. The use of machine learning methods yields both population-level effects and pinpoints the sub-groups of patients the program benefits the most. These methods exploit the potential of routine observational patient data within complex healthcare systems to inform decision-makers.

In chapter 4, we proposed a robust variable importance measure based on TMLE and applied it in estimating the transmission effects of mother's eating behavior on the next generation. To start with, we proposed a parameter within a non-parametric model that measures the importance of each variable as the amount of attribution of that variable towards changes in the mean outcome. The proposed parameter provides comparable results across continuous and categorical variables, and provides clinically meaningful interpretations on how changing the variable changes the outcome. To estimate such a parameter, we utilized an ensemble machine learning model as an initial estimator (Super Learner), and updated the estimator via target learning. This updated estimator, known as the Non-Parametric Variable Importance (NPVI) estimator, is not only independent of arbitrary model specifications, but also asymptotically linear (locally efficient) for which robust asymptotic inference is available. The linear coefficient can also be seen as an estimator of the proposed parameter under parametric model assumptions. We implemented both the linear coefficient and

the NPVI estimator to determine the influence of mother's and midlife eating behavior on child's body mass index (BMI) based on empirical data. Our results showed that if the mother has higher levels of drive for thinness, body dissatisfaction, bulimia, or interoceptive awareness, her child tend to have a higher level of BMI, adjusted for mother's adulthood stress, early-year socioeconomic status, health status and child's age and sex. We compared the performance of four estimators (with and without linear model specifications, and with and without target learning updates) under linear and non-linear simulation settings, and showed the robustness of the NPVI estimator. Our method for variable importance estimation can serve as an alternative to more standard variable importance procedures that lack both broad clinical interpretability and mechanisms for accurate statistical inference in the context of data-adaptive estimation.

# Chapter 2

# Population size estimation based on capture recapture and estimation reliability evaluation

## 2.1 Introduction

Epidemiologists use surveillance networks to monitor trends in disease frequency. When multiple surveillance components or surveys gather data on the same underlying population (such as those diagnosed with a particular disease over a particular time period), a variety of methods (capture-recapture designs [8], distance sampling [9], multiple observers [10], etc.) may be used to better estimate the disease occurrence in the population. Capture-recapture models are widely used for estimating the size of partially observed populations, usually assuming that individuals do not enter or leave the population between sample collections [11, 12]. These models have been widely applied to epidemiological data [13].

Due to the unobservability of outcomes for individuals not captured by any survey, additional identifying assumptions have to be made in capture-recapture problems. In two-sample scenarios, a common identification assumption is independence between the two samples (i.e. that capture in one survey does not change the probability of capture in the other survey). The estimator of population size based on this assumption is known as the Lincoln-Petersen estimator [14]. However, the independence assumption is often violated in empirical studies. In problems involving three or more surveys, it is common to assume that the highest order interaction term in log-linear or logistic model equals zero (i.e., that correlations between survey captures can be described by lower-order interactions alone), an assumption that is very difficult to interpret or empirically verify [11]. An additional challenge to capture-recapture estimators is the curse of dimensionality in the finite sample case, whereby the absence of one or more capture patterns from the observed sample leads to undefined interaction terms. Traditionally, a common approach to selecting among alternative capture-recapture models is to perform model selection based on Akaike's Information Criterion (AIC) [15], Bayesian

Information Criterion (BIC) [16], or Draper's version of the Bayesian Information Criterion [17]. However, this approach is known to have limited reliability in the presence of violations of its identifying assumptions [11, 18, 19]. Das and Kennedy made contributions on a doubly robust method under a specific identification assumption that two lists are conditionally independent given measured covariate information [20].

In this chapter, we propose a generalizable framework for estimating population size with as few model assumptions as possible. This framework can be adapted to posit various identification assumptions, including independence between any pair of surveys or absence of highest-order interactions, and can be applied to linear and nonlinear constraints. In high dimensional settings with finite samples, we use machine learning methods to smooth over unobserved capture patterns and apply targeted maximum likelihood estimation (TMLE) [2] updates to reduce estimation bias. Previous work has shown the vulnerability of the existing estimators with violations of identification assumptions [21, 22], and we further show the significant impact of the misspecified identification assumption on the estimation results for all existing and proposed estimators.

## Chapter outline

In this chapter, we start with the statistical formulation of the estimation problem. In section 2.2, we define the framework of our estimators under linear and non-linear constraints. Specifically, we develop the estimators under each of the following identification assumptions: no K-way interaction in a linear model, independence among samples, conditional independence among samples, and no K-way interaction in a log-linear model. In section 2.3, we derive the efficient estimator of the target parameter, and the statistical inference. In section 2.4 we provide the targeted learning updates for the smoothed estimators under the no K-way interaction log-linear model. In section 2.5, we illustrate the performance of existing and proposed estimators in various situations, including high-dimensional finite sample settings. In section 2.6, we show the performance of the estimators under violations of various identification assumptions. In section 4.2, we apply the estimators to surveillance data on a parasitic infectious disease in a region in southwestern China [23]. In section 2.8, we summarise the characteristics of the proposed estimators and state the main findings of the chapter.

## 2.2 Statistical formulation of estimation problem

### Defining the data and its probability distribution

We define the capture-recapture experiments in the following manner. One takes a first random sample from the population of size $n_1$, records identifying characteristics of the individuals captured and an indicator of their capture, then repeats the process a total of $K$ times, resulting in $K$ samples of size $n_1, \ldots, n_K$.

Each individual $i$ in the population, whether captured or not, defines a capture history as a vector $B_i^* = (B_i^*(1), \ldots, B_i^*(K))$, where $B_i^*(k)$ denotes the indicator that this individual $i$ is captured by sample $k$. We assume the capture history vectors $B_i^*$ of all $i = 1, \ldots, N$ individuals independently and identically follow a common probability distribution, denoted by $P_{B^*}$. Thus $P_{B^*}$ is defined on $2^K$ possible vectors of dimension $K$.

Note that, for the individuals contained in our observed sample of size $n = \sum_{k=1}^{K} n_k$, we actually observe the realization of $B^*$. For any individual $i$ that is never captured, we know that $B_i^* = 0$, where $0 = (0, \ldots, 0)$ is the $K$ dimensional vector in which each component equals 0. However, for these $N-n$ individuals we do not know the identity of these individuals and we also do not know how many there are (i.e., we do not know $N - n$).

Therefore, we can conclude that our observed data set $B_1, \ldots, B_n$ are $n$ independent and identically distributed draws from the conditional distribution of $B^*$, given $B^* \neq 0$. Let's denote this true probability distribution with $P_0$ and its corresponding random variable with $B$.

So we can conclude that under our assumptions we have that $B_1, \ldots, B_n \sim_{iid} P_0$, where $P_0(b) = P_{B^*,0}(b \mid B \neq 0)$ for all $b \in \{0,1\}^K$. Since the probability distribution $P$ of $B$ is implied by the distribution $P^*$ of $B^*$ we also use the notation $P = P_{P^*}$ to stress this parameterization.

### Full-data model, target quantity

Let $M^F$ be a model for the underlying distribution $P_{B^*,0}$. The full-data target parameter $\Psi^F : M^F \to I\!R$ is defined as

$$\Psi^F(P_{B^*}) = P_{B^*}(B \neq 0).$$

In other words, we want to know the proportion of $N$ that on average will not be caught by the combined sample. An estimator $\psi_n^F$ of this $\psi_0^F = \Psi^F(P_{B^*,0})$ immediately translates into an estimator of the desired size $N$ of the population of interest: $N_n = \frac{n}{\psi_n^F}$. We will focus on full-data models defined by a single constraint, where we distinguish between linear constraints defined by a function $f$ and a non-linear constraint defined by a function $\Phi$. Specifically, for a given function $f$, we define the full-data model

$$M_f^F = \left\{ P_{B^*} : \sum_b f(b) P_{B^*}(b) = 0, 0 < P^*(0) < 1 \right\}.$$

We will require that $f$ satisfies the following condition:

$$f(b=0) \neq 0. \tag{2.1}$$

In other words, $M_f^F$ contains all probability distributions of $B^*$ for which $E_{P_{B^*}} f(B^*) = 0$. An example of interest is:

$$f_I(b) = (-1)^{K+\sum_{k=1}^K b_k}.$$

In this case, $E_0 f_I(B^*) = 0$ is equivalent with

$$\sum_b (-1)^{K+\sum_{k=1}^K b_k} P_{B^*,0}(b) = 0. \tag{2.2}$$

We note the left-hand side represents a $K$-th way interaction term $\alpha_1$ in the saturated model

$$P_{B^*}(b) = \alpha_0 + \sum_{b' \neq 0} \alpha_{b'} \prod_{j:b'_j=1} b_j.$$

For example, if $K = 2$, then the latter model states:

$$P_{B^*}(b_1, b_2) = \alpha_{00} + \alpha_{10} b_1 + \alpha_{01} b_2 + \alpha_{11} b_1 b_2,$$

and the constraint $E f_I(B^*) = 0$ states that $\alpha_{11} = 0$.

One might also use a log-link in this saturated model:

$$\log P_{B^*}(b) = a_0 + \sum_{b' \neq 0} a_{b'} \prod_{j:b'_j=1} b_j.$$

In this case, $a_1$ is the $K$-way interaction term in this log-linear model, and we now have

$$a_1 \equiv \sum_b (-1)^{K+\sum_{k=1}^K b_k} \log P_{B^*,0}(b) = 0.$$

So, assuming $a_1 = 0$ corresponds with assuming

$$0 = \sum_b (-1)^{K+\sum_{k=1}^K b_k} \log P_{B^*,0}(b).$$

This is an example of a non-linear constraint $\Phi_I(P^*) = 0$, where

$$\Phi_I(P^*) \equiv \sum_b (-1)^{1+\sum_{k=1}^K b_k} \log P_{B^*}(b). \tag{2.3}$$

We will also consider general non-linear constraints defined by such a function $\Phi$, so that for that purpose one can keep this example $\Phi_I$ in mind. As we will see the choice of this constraint has quite dramatic implications on the resulting statistical target parameter/estimand and thereby on the resulting estimator. As one can already tell from the definition of $\Phi_I$,

$\Phi_I$ is not even defined if there are some $b$ for which $P_{B^*}(b) = 0$, so that also an NPMLE of $\Psi^F_{\Phi_I}(P^*_0)$ will be ill defined in the case that the empirical distribution $P_n(B = b) = 0$ for some $b \neq 0$. As we will see the parameter $\Psi^F_{f_I}(P^*_0)$ is very well estimated by the MLE, even when $K$ is very large relative to $n$, but for $\Psi^F_{\Phi_I}(P^*_0)$ we would need a so called TMLE, incorporating machine learning [2].

Since it is hard to believe that the $\Phi_I$ constraint is more realistic than the $f$-constraint in real applications, it appears that, without a good reason to prefer $\Phi_I$, the $f$-constraint is far superior. However, it also raises alarm bells that the choice of constraint, if wrong, can result in dramatically different statistical output, so that one should really try to make sure that the constraint that is chosen is known to hold by design.

Another example of a non-linear constraint is the independence between two samples, i.e., that

$$P^*(B^*(1:2) = (0,0)) = P^*(B^*(1) = 0)P^*(B^*(2) = 0),$$

i.e., that the binary indicators $B^*(1)$ and $B^*(2)$ are independent, but the remaining components can depend on each other and depend on $B^*(1), B^*(2)$. This corresponds with assuming $\Phi_{II}(P^*) = 0$, where

$$\Phi_{II}(P^*) \equiv \sum_b I(b(1:2) = (0,0))P^*(b) - \sum_{b_1,b_2} I(b_1(1) = b_2(2) = 0)P^*(b_1)P^*(b_2). \tag{2.4}$$

A third non-linear constraint example is conditional independence between two samples, given the others. Suppose we have $K$ samples in total, and the distribution of the $j^{th}$ sample $B_j$ is independent of the $m^{th}$ sample $B_m$ given all other samples (there's no time ordering of $j, m$). Then we can derive the target parameter and efficient influence curve for this conditional independence constraint. The conditional independence constraint $\Phi_{CI} = 0$ is defined as

$$\begin{aligned}\Phi_{CI,(j,m)} = P^*(B_j = 1|B_1 = b_1, \cdots, B_m = 0, \cdots, B_K = b_k)\\ -P^*(B_j = 1|B_1 = b_1, \cdots, B_m = 1, \cdots, B_K = b_k), b_t = 0, 1, t = 1, \cdots, K.\end{aligned}$$

Because we must have the term $P^*(0, 0, \cdots, 0)$ in the constraint to successfully identify the parameter of interest, the constraint $\Phi_{CI} = 0$ is sufficient. Thus the equation above can be presented as

$$\begin{aligned}\Phi_{CI,(j,m)} = P^*(B_j = 1|B_1 = 0, \cdots, B_m = 0, \cdots, B_K = 0)\\ -P^*(B_j = 1|B_1 = 0, \cdots, B_m = 1, \cdots, B_K = 0).\end{aligned} \tag{2.5}$$

This is because for all the combinations of

$$\{B_1 = b_1, \cdots, B_{m-1} = b_{m-1}, B_{m+1} = b_{m+1}, \cdots, B_K = b_K\}, \forall b_t \in \{0, 1\}, t = 1, \cdots, K,$$

only the situation of $\{B_1 = 0, \cdots, B_{m-1} = 0, B_{m+1} = 0, \cdots, B_K = 0\}$ can generate the required term $P^*(0, 0, \cdots, 0)$ in the constraint.

## Identifiability from probability distribution of data

Given such a full-data model with one constraint defined by $f$ or $\Phi$, one will need to establish that for all $P^* \in M_f^F$, we have (say we use $f$)

$$\Psi^F(P^*) = \Psi_f(P).$$

for a known $\Psi_f : M \to (0,1)$ and $P = P_{P^*}$, where the statistical model for $P_0$ is defined as

$$M_f = \{P_{P_{B^*}} : P_{B^*} \in M_f^F\}.$$

Let's first study this identifiability problem in the special case of our full data models defined by the linear constraint $P_0^* f = 0$ with $f(0) \neq 0$ (equation 2.2). It will show that we have identifiability of the whole $P_{B^*}$ from $P = P_{P_{B^*}}$. Firstly, we note that $P_{P^*}(b) = P^*(b)/\psi^F$ for all $b \neq 0$. Thus, $P^*(b) = \psi^F P(b)$ for all $b \neq 0$. We also have $P^*(0) = 1 - \psi^F$, so that $\sum_b f(b)P^*(b) = 0$ yields:

$$\begin{aligned} 0 &= \sum_b f(b)P^*(b) = \sum_{b\neq 0} f(b)\psi^F P(b) + f(0)P^*(0) \\ &= \psi^F \sum_{b\neq 0} f(b)P(b) + f(0)(1-\psi^F). \end{aligned}$$

We can now solve for $\psi^F$:

$$\psi^F = \frac{f(0)}{f(0) - \sum_{b\neq 0} f(b)P(b)}.$$

At first sight, one might wonder if this solution is in the range $(0,1)$. Working backwards from the right-hand side, i.e., using $f(0)P^*(0) + \sum_{b\neq 0} f(b)P^*(b) = 0$ and $P^*(b) = P(b)/\psi^F$, it indeed follows that the denominator equals $f(0) + f(0)(1-\psi^F)/\psi^F$, so that the right-hand side is indeed in $(0,1)$. Thus, we can conclude that:

$$\Psi^F(P^*) = \Psi_f(P) \equiv \frac{f(0)}{f(0) - Pf}, \tag{2.6}$$

where we use the notation $Pf = \int f(b)dP(b)$ for the expectation operator.

Suppose now that our the one-dimensional constraint in the full-data model is defined in general by $\Phi(P^*) = 0$ for some function $\Phi : M_\Phi^F \to I\!R$. The full data model is now defined by $M_\Phi^F = \{P^* : \Phi(P^*) = 0, 0 < P^*(0) < 1\}$, and the corresponding observed data model can be denoted with $M_\Phi$. To establish the identifiability, we still use $P^*(b) = \psi^F P(b)$ for all $b \neq 0$. We also still have $P^*(0) = 1 - \psi^F$. Define $P_{P,\psi}^*$ as $P_{P,\psi}^*(0) = 1 - \psi$ and $P_{P,\psi}^*(b) = \psi P(b)$ for $b \neq 0$. The constraint $\Phi(P^*) = 0$ now yields the equation

$$\Phi(P_{P,\psi^F}^*) = 0 \text{ in } \psi^F$$

for a given $P = P_{P^*}$.

Consider now our particular example $\Phi_I$. Note that

$$\Phi_I(P^*_{P,\psi}) = \sum_{b\neq 0}(-1)^{K+\sum_k b_k}\log\{\psi P(b)\} + (-1)^K \log(1-\psi).$$

We need to solve this equation in $\psi$ for the given $P$. For $K$ is odd, we obtain

$$\Psi_I(P) = \frac{1}{1+\exp\left(\sum_{b\neq 0} f_I(b)\log P(b)\right)},$$

and for $K$ even we obtain $\log(1-\psi)/\psi) = \sum_{b\neq 0}\log P(b)$ and thus

$$\Psi_I(P) = \frac{1}{1+\exp\left(-\sum_{b\neq 0} f_I(b)\log P(b)\right)}.$$

So, in general, this solution can be represented as:

$$\Psi_I(P) = \frac{1}{1+\exp\left((-1)^{K+1}\sum_{b\neq 0} f_I(b)\log P(b)\right)}. \tag{2.7}$$

This solution $\Psi(P)$ only exists if $P(b) > 0$ for all $b \neq 0$. In particular, a plug-in estimator $\Psi(P_n)$ based on the empirical distribution function $P_n$ would not be defined when $P_n(b) = 0$ for some cells $b \in \{0,1\}^K$.

For general $\Phi$, one needs to assume that this one-dimensional equation $H(\psi^F, P) \equiv \Phi(P^*_{P,\psi}) = 0$ in $\psi$, for a given $P \in M$, always has a unique solution, which then proves the desired identifiability of $\psi^F(P^*)$ from $P_{P^*}$ for any $P^* \in M^F_\Phi$. This solution is now denoted with $\Psi_\Phi(P)$. So, in this case, $\Psi_\Phi : M_\Phi \to I\!R$ is defined implicitly by $H(\Psi_\Phi(P), P) = 0$. If $\Phi$ is one-dimensional, one will still have that $M_\Phi$ is nonparametric.

Let's now consider the $\Phi_{II}$ constraint which assumes that $B^*(1)$ is independent of $B^*(2)$. Again, using $P^*(b) = P(b)\psi$ for $b \neq 0$ and $P^*(0) = (1-\psi)$, the equation $\Phi_{II}(P^*) = 0$ yields the following quadratic equation in $\psi$:

$$a_{II}(P)\psi^2 + b_{II}(P)\psi = 0,$$

where

$$\begin{aligned}
a_{II}(P) &= -\sum_{b_1\neq 0, b_2\neq 0} I(b_1(1) = b_2(2) = 0)P(b_1)P(b_2) \\
&\quad + \sum_{b_2\neq 0} I(b_2(2) = 0)P(b_2) + \sum_{b_1\neq 0} I(b_1(1) = 0)P(b_1) - 1 \\
b_{II}(P) &= \sum_{b\neq 0} I(b(1:2) = (0,0))p(b) - \sum_{b_2\neq 0} I(b_2(2) = 0)P(b_2) \\
&\quad - \sum_{b_1\neq 0} I(b_1(1) = 0)P(b_1) + 1.
\end{aligned}$$

Since $\psi \neq 0$, this yields the equation $a_{II}(P)\psi + b_{II}(P) = 0$ and thus

$$\Psi_{II}(P) = \frac{-b_{II}(P)}{a_{II}(P)}.$$

A more helpful way this parameter can be represented is given by:

$$\Psi_{II}(P) = \frac{1 - P(B(1) = 0) - P(B(2) = 0) + P(B(1:2) = (0,0))}{1 - P(B(1) = 0) - P(B(2) = 0) + P(B(1) = 0)P(B(2) = 0)}. \tag{2.8}$$

This identifiability result relies on $\Psi_{II}(P) \in (0,1)$, i..e, that $P(B(1) = B(2) = 0) < P(B(1) = 0)P(B(2) = 0)$. In particular, we need $0 < P(B(1) = 0) < 1$ and $0 < P(B(2) = 0) < 1$ and thereby that each of the three cells $(1,0),(0,1),(1,1)$ has positive probability under the bivariate marginal distribution of $B(1), B(2)$ under $P$. It follows trivially that the inequality holds for $K = 2$ since in that case $P(B(1) = B(2) = 0) = 0$. It will have to be verified if this inequality constraint always holds for $K > 2$ as well, or that this is an actual assumption in the statistical model. Since it would be an inequality constraint in the statistical model, it would not affect the tangent space and thereby the efficient influence function for $\Psi_{II} : M \to I\!R$ presented below, but the MLE would now involve maximizing the likelihood under this constraint so that the resulting MLE of $P_0$ satisfies this constraint.

Similarly, for the conditional independence assumption, the parameter $\Psi_{CI}$ can be derived as

$$\Psi_{CI} = \frac{P(B_m = 1, B_j = 1, 0, ..., 0)}{C_0(P)}, \tag{2.9}$$

where

$$\begin{aligned} C_0(P) &= P(B_m = 1, B_j = 1, 0, ..., 0) \\ &\quad + P(B_m = 0, B_j = 1, 0, ..., 0)P(B_m = 1, B_j = 0, 0, ..., 0). \end{aligned} \tag{2.10}$$

Details on this derivation are presented in the appendix 2.9.

One could also define a model by a multivariate $\Phi : M^F \to I\!R^d$. In this case, the full data model is restricted by $d$ constraints $\Phi(P^*) = 0$ and the observed data model will not be saturated anymore. Let's consider the example in which we assume that $(B^*(1), \ldots, B^*(K))$ are $K$ independent Bernoulli random variables. In this case, we assume that $P^*(B^* = b) = \prod_{k=1}^K P^*(B^*(k) = b(k))$ for all possible $b \in \{0,1\}^K$. This can also be defined as stating that for each 2 components $B^*(j_1), B^*(j_2)$, we have that these two Bernoulli's are independent. Let $\Phi_{,II,j_1,j_2}(P)$ be the constraint defined as in (2.4) but with $B(1)$ and $B(2)$ replaced by $B(j_1)$ and $B(j_2)$, respectively. Then, we can define $\Phi_{III}(P) = (\Phi_{II,j_1,j_2}(P) : (j_1,j_2) \in \{1,\ldots,K\}^2, j_1 \neq j_2)$, a vector of dimension $K(K-1)/2$. This defines now the model $M^F_{III} = \{P^* : \Phi_{III}(P^*) = 0\}$, which assumes that all components of $B^*$ are independent. We will also work out the MLE and efficient influence curve of $\psi_0^F$ for this restricted statistical model.

## Statistical Model and target parameter

We have now defined the statistical estimation problem for linear and non-linear constraints. For linear constraints defined by a function $f$, we observe $B_1, \ldots, B_n \sim P_0 \in M_f = \{P_{P^*} : P^* \in M^F\}$, $M^F = \{P^* : P^* f = 0, 0 < P^*(0) < 1\}$, and our statistical target parameter is given by $\Psi_f : M \to I\!R$, where

$$\Psi_f(P) = \frac{f(0)}{f(0) - Pf}. \tag{2.11}$$

The statistical target parameter satisfies that $\Psi^F(P^*) = P^*(B^* \neq 0) = \Psi_f(P_{P^*})$ for all $P^* \in M^F$.

Since the full-data model only includes a single constraint, it follows that $M_f$ consists of all possible probability distributions of $B$ on $\{b : b \neq 0\}$, so that it is a nonparametric/saturated model.

Similarly, we can state the statistical model and target parameter for our examples using non-linear one-dimensional constraints $\Phi$.

In general, we have a statistical model $M = \{P_{P^*} : P^* \in M^F\}$ for some full data model $M^F$ for the distribution of $B^*$, and, we would be given a particular mapping $\Psi : M \to I\!R$, satisfying $\Psi^F(P^*) = \Psi(P_{P^*})$ for all $P^* \in M^F$. In this case, our goal is to estimate $\Psi(P_0)$ based on knowing that $P_0 \in M$.

## Efficient influence curve of target parameter

An estimator of $\Psi$ is efficient if and only if it is asymptotically linear with influence curve equal to canonical gradient of pathwise derivative of $\Psi$. Therefore it is important to determine this canonical gradient. It teaches us how to construct an efficient estimator of $\Psi(P)$ in model $M^F$. In addition, it provides us with Wald type confidence intervals based on an efficient estimator. As we will see for most of our estimation problems $\Psi$ with a nonparametric model, $P$ and $\Psi$ can be estimated with the empirical measure $P_n$ and $\Psi(P_n)$. However for constraint $\Phi_I$, an NPMLE is typically not defined due to empty cells, so that smoothing and bias reduction is needed.

Let $\Psi_{1f}(P) = \sum_{b \neq 0} f(b)P(b)$ so that $\Psi_f(P) = \frac{f(0)}{f(0) - \Psi_{1f}(P)}$. Note that $\Psi_{1f}(P) = Pf$ is simply the expectation of $f(B)$ w.r.t its distribution. $\Psi_{1f} : M_f \to I\!R$ is pathwise differentiable parameter at any $P \in M$ with canonical gradient/efficient influence curve given by:

$$D^*_{1f}(P)(B) = f(B) - \Psi_{1f}(P).$$

By the delta-method the efficient influence curve of $\Psi_f$ at $P$ is given by:

$$D^*_f(P)(B) = \frac{f(0)}{\{f(0) - \Psi_{1f}\}^2} D^*_{1f}(P)(B). \tag{2.12}$$

In general, if $\Psi : M \to I\!R$ is the target parameter and the model $M$ is nonparametric, then the efficient influence curve $D^*(P)$ is given by $D^*(P) = d\Psi(P)(P_{n=1} - P)$, where

$$d\Psi(P)(h) = \frac{d}{d\epsilon}\Psi(P + \epsilon h)\bigg|_{\epsilon=0}$$

is the Gateaux derivative in the direction $h$, and $P_{n=1}$ is the empirical distribution for a sample of size one $\{B\}$, putting all its mass on $B$ [24].

Consider now the model $M_\Phi$ for a general univariate constraint function $\Phi : M_{NP}^F \to I\!R$ that maps any possible $P^*$ into a real number. Recall that $\Psi_\Phi : M_\Phi \to I\!R$ is now defined implicitly by $\Phi(P^*_{P,\psi}) = 0$. For notational convenience, in this particular paragraph we suppress the dependence of $\Psi$ on $\Phi$. The implicit function theorem implies that the general form of efficient influence curve is given by:

$$D^*_\Phi(P)(B) = -\left\{\frac{d}{d\psi}\Phi(P^*_{P,\psi})\right\}^{-1}\frac{d}{dP}\Phi(P^*_{P,\psi})(P_{n=1} - P).$$

where the latter derivative is the directional derivative of $P \to \Phi(P^*_{P,\psi})$ in the direction $P_{n=1} - P$.

Note that

$$\frac{d}{d\psi}\Phi(P^*_{P,\psi}) = d\Phi(P^*_{P,\psi})\frac{d}{d\psi}P^*_{P,\psi}.$$

We now use that $P^*_{P,\psi}(b) = P(b)I(b \neq 0) + (1 - \psi)I(b = 0)$. So we obtain:

$$\frac{d}{d\psi}\Phi(P^*_{P,\psi}) = d\Phi(P^*_{P,\psi})(-1I_0),$$

where $I_0(b)$ is the function in $b$ that equals 1 if $b = 0$ and zero otherwise, and

$$d\Phi(P^{*0})(h) \equiv \frac{d}{d\epsilon}\Phi(P^{*0} + \epsilon h)\bigg|_{\epsilon=0}$$

is the directional/Gateaux derivative of $\Phi$ in the direction $h$. We also have:

$$\frac{d}{dP}\Phi(P^*_{P,\psi})(P_{n=1} - P) = d\Phi(P^*_{P,\psi})\frac{d}{dP}P^*_{P,\psi}(P_{n=1} - P).$$

We have $\frac{d}{dP}P^*_{P,\psi}(h) = I_{0^c}h$, where $I_{0^c}(b)$ is the function in $b$ that equals zero if $b = 0$ and equals 1 otherwise. So we obtain:

$$\frac{d}{dP}\Phi(P^*_{P,\psi})(P_{n=1} - P) = d\Phi(P^*_{P,\psi})(I_{0^c}(P_{n=1} - P)).$$

We conclude that

$$D^*_\Phi(P) = -\{d\Phi(P^*_{P,\psi})(-1I_0)\}^{-1}d\Phi(P^*_{P,\psi})(I_{0^c}(P_{n=1} - P)).$$

Let's now consider our special example $\Phi_I$. In this case, the statistical target parameter $\Psi_I : M \to I\!R$ is given by (2.7). It is straightforward to show that the directional derivative of $\Psi$ at $P = (P(b) : b)$ in direction $h = (h(b) : b)$ is given by:

$$d\Psi_I(P)(h) = (-1)^K \Psi_I(P)(1 - \Psi_I(P)) \sum_{b \neq 0} \frac{f_I(b)}{P(b)} h(b).$$

As one would have predicted from the definition of $\Psi_I$, this directional derivative is only bounded if $P(b) > 0$ for all $b \neq 0$. The efficient influence curve is thus given by $d\Psi_I(P)(P_{n=1} - P)$:

$$\begin{aligned} D^*_{\Phi_I}(P) &= (-1)^K \Psi_I(P)(1 - \Psi_I(P)) \sum_{b \neq 0} \frac{f_I(b)}{P(b)} \{I_B(b) - P(b)\} \\ &= (-1)^K \Psi_I(P)(1 - \Psi_I(P)) \left\{ \frac{f_I(B)}{P(B)} + f_I(0) \right\}, \end{aligned} \tag{2.13}$$

where we use that $\sum_{b \neq 0} f_I(b) = -f_I(0)$ so that indeed the expectation of $D^*_{\Phi_I}(P)$ equals zero (under $P$).

The efficient influence function for $\Psi_{II} : M \to I\!R$ (2.8), corresponding with the constraint $\Phi_{II}(P^*) = 0$, is the influence curve of the empirical plug-in estimator $\Psi_{II}(P_n)$, and can thus be derived from the delta-method:

$$\begin{aligned} D^*_{\Phi_{II}}(P) &= C_2(P)\{I(B(1) = 0) - P(B(1) = 0)\} \\ &\quad + C_3(P)\{I(B(2) = 0) - P(B(2) = 0)\} \\ &\quad + C_4(P)\{I(B(1) = B(2) = 0) - P(B(1:2) = 0)\}, \end{aligned} \tag{2.14}$$

where

$$\begin{aligned} C_2(P) &= \frac{1 - P(B(1)=0) - P(B(2)=0) - P(B(1)=B(2)=0)}{(1 - P(B(1)=0) - P(B(2)=0) - P(B(1)=0)P(B(2)=0))^2} P(B(2) = 0) \\ C_3(P) &= \frac{1 - P(B(1)=0) - P(B(2)=0) - P(B(1)=B(2)=0)}{(1 - P(B(1)=0) - P(B(2)=0) - P(B(1)=0)P(B(2)=0))^2} P(B(1) = 0) \\ C_4(P) &= -\frac{1}{1 - P(B(1)=0) - P(B(2)=0) - P(B(1)=0)P(B(2)=0)}. \end{aligned}$$

Similarly, we can derive the efficient influence curve for conditional independence constraint $\Phi_{CI}$ and parameter $\Psi_{CI}$ under this constraint as

$$D^*_{\Phi_{CI}}(P) = \frac{1}{C_5(P)} (C_6(P) - C_7(P) - C_8(P)). \tag{2.15}$$

where

$$\begin{aligned}
C_5(P) &= -\sum_{b_1\neq 0, b_2\neq 0} I(b_1(1)=b_2(2)=0)P(b_1)P(b_2).\\
C_6(P) &= P(B_m=0,B_j=1,0,...,0)P(B_m=0,B_j=0,0,...,0)\\
&\quad [\mathbb{I}(B_m=1,B_j=1,0,...,0)-P(B_m=1,B_j=1,0,...,0)].\\
C_7(P) &= P(B_m=1,B_j=0,0,...,0)P(B_m=1,B_j=1,0,...,0)\\
&\quad [\mathbb{I}(B_m=0,B_j=1,0,...,0)-P(B_m=0,B_j=1,0,...,0)].\\
C_8(P) &= P(B_m=0,B_j=1,0,...,0)P(B_m=1,B_j=1,0,...,0)\\
&\quad [\mathbb{I}(B_m=1,B_j=0,0,...,0)-P(B_m=1,B_j=0,0,...,0)].
\end{aligned}$$

The details for deriving the influence curve $D^*_{\Phi_{CI}}(P)$ can be found in appendix, section 2.9.

Finally, consider a non-saturated model $M$ implied by a multidimensional constraint function $\Phi$. In this case, the above $D^*_\Phi(P)$ is still a gradient of the pathwise derivative, where division by a vector $x$ is now defined as $1/x = (1/x_j : j)$ component wise. However, this is now not equal to the canonical gradient. We can now determine the tangent space of the model $M$ and project $D^*_\Phi(P)$ onto the tangent space at $P$, which then yields the actual efficient influence curve. We can demonstrate this for the example defined by the multidimensional constraint $\Phi_{III}$ using the general result in appendix 2.9.

## 2.3 Efficient estimator of target parameter, and statistical inference

Estimation of $\Psi_f(P_0)$ based on statistical model $M_f$ and data $B_1, \ldots, B_n \sim_{iid} P_0$ is trivial since $\Psi_{1f}(P_0) = P_0 f$ is just a mean of $f(B)$. In other words, we estimate $\Psi_{1f}(P_0)$ with the NPMLE

$$\Psi_{1f}(P_n) = P_n f = \frac{1}{n}\sum_{i=1}^{n} f(B_i),$$

where $P_n$ is the empirical distribution of $B_1, \ldots, B_n$, and, similarly, we estimate $\Psi_f(P_0)$ with its NPMLE

$$\Psi_f(P_n) = \frac{f(0)}{f(0) - \Psi_{1f}(P_n)} = \frac{f(0)}{f(0) - P_n f}.$$

This estimator is asymptotically linear at $P_0$ with influence curve $D^*_\Phi(P_0)$ under no further conditions. As a consequence, a valid asymptotic 95% confidence interval is given by:

$$\Psi_f(P_n) \pm q(0.975)\sigma_n/\sqrt{n},$$

where

$$\sigma_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\{D^*_f(P_n)(B_i)\}^2,$$

and $q(0.975)$ is the 0.975 quantile value of standard normal distribution. If the general constraint $\Phi : M \to I\!R$ is differentiable so that $\Phi(P_n)$ is an asymptotically linear estimator of $\Phi(P_0)$ [25], then, we can estimate $\Psi_\Phi(P_0)$ with the NPMLE $\Psi_\Phi(P_n)$. Again, under no further conditions, $\Psi_\Phi(P_n)$ is asymptotically linear with influence curve $D^*_\Phi(P_0)$, and an asymptotically valid confidence interval is obtained as above. The estimator $\Psi_f(P_n)$ is always well behaved, even when $n$ is relatively large relative to $K$. For a general $\Phi$, this will very much depend on the the precise dependence on $P$ of $\Phi(P)$. In general, if $\Phi(P_n)$ starts suffering from the dimensionality of the model (i.e., empty cells make the estimator erratic), then one should expect that $\Psi_\Phi(P_n)$ will suffer accordingly, even though it will still be asymptotically efficient. In these latter cases, we propose to use targeted maximum likelihood estimation, which targets data-adaptive machine learning fits towards optimal variance-bias trade-off for the target parameter.

Let's consider the $\Phi_I$-example which assumes that the $K$-way interaction on the log-scale equals zero. In this case the target parameter $\Psi_I(P)$ is defined by (2.7), which shows that the NPMLE $\Psi_I(P_n)$ is not defined when $P_n(b) = 0$ for some $b \in \{0,1\}^K$. So in this case, the MLE suffers immensely from the curse of dimensionality and can thus not be used when the number $K$ of samples is such that the sample size $n$ is of the same order as $2^K$. In this context, we discuss estimation based on TMLE in the next section.

$\Psi_{II}(P_0)$ can be estimated with the plug-in empirical estimator which is the NPMLE. So, this estimator only relies on positive probability on $(0,1), (1,0)$ and $(1,1)$ under the bivariate distribution of $B(1), B(2)$ under $P$. We also note that this efficient estimator of

$\psi_{II,0}$ does only use the data on the first two samples. Thus, the best estimator based upon this particular constraint $\Phi_{II}$ ignores the data on all the other samples. More assumptions will be needed, such as the model that assumes that all components of $B^*$ are independent, in order to create a statistical model that is able to also incorporate the data with the other patterns.

Constrained models based on multivariate $\Phi$: For these models, if the NPMLE would behave well for the larger model in which just one of the $\Phi$ constraints is used, then it will behave well under more constraints. If on the other hand, this NPMLE suffers from curse of dimensionality, it might be the case that the MLE behaves better due to the additional constraints, but generally speaking that is a lot to hope for (except if $\Phi$ is really high dimensional relative to $2^K$). To construct an asymptotically efficient estimator one can thus use the MLE $\Psi_\Phi(\tilde{P}_n)$ where now $\tilde{P}_n$ is the MLE over the actual model $M_\Phi$. In the special case that the behavior of this MLE $\Psi_\Phi(\tilde{P}_n)$ suffers from the curse of dimensionality, we recommend the TMLE below instead, which will be worked out for the example $\Phi_I$.

## 2.4 Targeted maximum likelihood estimation when the (NP)MLE of target parameter suffers from curse of dimensionality

Consider the statistical target parameter $\Psi_I : M_{\Phi_I} \rightarrow I\!R$ defined by (2.7), whose efficient influence curve is given by (2.13). Let $P_n^0$ be an initial estimator of the true distribution $P_0$ of $B$ in the nonparametric statistical model $M_{\Phi_I}$ that consists of all possible distributions of $B = (B(1), \ldots, B(K))$. An ideal initial estimator for $P$ would be consistent and identifiable when empty cells exist. One could use, for example, an undersmoothed lasso estimator constructed in algorithm 1 as the initial estimator [26]. In algorithm 1, we establish the empirical criterion by which the level of undersmoothing may be chosen to appropriately satisfy the conditions required of an efficient plug-in estimator. In particular, we require that the minimum of the empirical mean of the selected basis functions is smaller than a constant times $n^{-\frac{1}{2}}$, which is not parameter specific. This condition essentially enforces the selection of the $L_1$-norm in the Lasso to be large enough so that the fit includes sparsely supported basis functions [26]. When the hyperparameter $\lambda$ equals zero, the undersmoothed lasso estimator is the same as the NPMLE plug-in estimator, and when $\lambda$ is not zero, the undersmoothed lasso estimator will smooth over the predicted probabilities and avoid predicting probabilities of exact zero when empty cells exist.

**Algorithm 1** Undersmoothed Lasso

The log-linear model can be expressed as:

$$\log E_{B^*}(b) = a_0 + \sum_{b \neq 0} a_b \prod_{j:b_j=1} b_j.$$

In this case, $b = (b_1, b_2, ..., b_K)$, and $b_j = 1$ if the subject is captured by sample $j, j = 1, 2, ..., K$. $E_{B^*}(b)$ is the count of observations in cell $b$. The $K$-way interaction term is $a_{1,1,...,1}$(K terms of 1), and the identification assumption is that $a_{1,1,...,1} = 0$.

Fit a lasso regression $M_{lasso}$ with all but the highest way interaction term as the independent variable, $E_{B^*}(b)$ (the count) as the dependent variable, and specify model family as Poisson. The regularization term $\lambda$ is chosen such that the absolute value of the empirical mean of the efficient influence function $P_n D^*_{\Phi_I}(P_n)(B_i) \leq T^* = \frac{\sigma_n}{\sqrt{n}}$, where $\sigma_n = \sqrt{\frac{1}{n}\sum_{i=1}^n \{D^*_{I,cv}(P_n)(B_i)\}^2}$. The probabilities $P_n(B_i), i = 1, ..., n$ used in computing $\sigma_n$ are estimated by the lasso regression with regularization term chosen by cross-validation. This algorithms will naturally avoid fits with $P_n(B_i) = 0$ for any $B_i$ since that would result in a log-likelihood equal to minus infinity.

**while** $P_n D^*_{\Phi_I}(P_n)(B_i) \geq T^*$ **do**

1. decrease $\lambda$;
2. predict count $\hat{E}_{B^*}(b)$ using $M_{lasso}$ and the new $\lambda$;
3. calculate predicted probability $\hat{P}_B(b)$ by dividing the count $\hat{E}_{B^*}(b)$ by the sum of all counts;
4. calculate $\hat{\Psi}_I(P) = \frac{1}{1+exp((-1)^K \sum_{b\neq 0} f_I(b) log \hat{P}_B(b))}$;
5. calculate $D^*_{\Phi_I}(P_n)(B_i) = \hat{\Psi}_I(P)(1 - \hat{\Psi}_I(P))[\frac{f_I(b)}{\hat{P}_B(b)} + f_I(0)]$;
6. update $P_n D^*_{\Phi_I}(P_n)(B_i) = \frac{1}{n}\sum_{i=1}^n D^*_{\Phi_I}(P_n)(B_i)$.

**end**

**Result:** Predicted probability $\hat{P}_B(b)$ for each cell

We denote $P_n^0 \equiv \hat{P}_B(b)$ as our initial estimator, the TMLE $P_n^*$ will update this initial estimator $P_n^0$ in such a way that $P_n D^*_{\Phi_I}(P_n^*) = 0$, allowing a rigorous analysis of the TMLE $\Psi_I(P_n^*)$ of $\Psi_I(P_0)$ as presented in algorithm 2.

Note that indeed this submodel 2.16 $\{P_\epsilon : \epsilon\} \subset M_{\Phi_I}$ through $P$ is a density for all $\epsilon$ and that its score is given by:

$$\frac{d}{d\epsilon} \log P_\epsilon \Big|_{\epsilon=0} = D^*_{\Phi_I}(P).$$

Recall that $\sum_{b\neq 0} f_I(b) = -f_I(0)$, so that indeed the expectation of its score equals zero: $E_P D^*_{\Phi_I}(P)(B) = 0$. This proves that if we select as loss function of $P$ the log-likelihood loss $L(P) = -\log P$, then the score $\frac{d}{d\epsilon} L(P_\epsilon)\big|_{\epsilon=0}$ spans the efficient influence function $D^*_{\Phi_I}(P)$. This property is needed to establish the asymptotic efficiency of the TMLE, with details in appendix section 2.9.

In accordance with general TMLE procedures [2], the TMLE updating process in algorithm 2 will be iterated till convergence at which point $\epsilon_n^m \approx 0$, or $P_n D^*_{\Phi_I}(P_n^m) = o_P(1/\sqrt{n})$. Let $P_n^* = \lim_m P_n^m$ be the probability distribution in the limit. The TMLE of $\Psi_I(P_0)$ is now defined as $\Psi_I(P_n^*)$. Due to the fact that the MLE $\epsilon_n^m$ solves its score equation and $\epsilon_n^m \approx 0$, it follows that this TMLE satisfies $P_n D^*_{\Phi_I}(P_n^*) = 0$ (or $o_P(1/\sqrt{n})$).

**Algorithm 2** Targeted maximum likelihood estimation(TMLE) update procedure

**Input**: Vector of predicted probability $\hat{P}_B(b)$; observed population size $n$;
**Procedure**:
1. Denote $P_n^0 = \hat{P}_B(b)$, calculate $\Psi_n(P_n^0)$ using equation 2.7, $D^*_{\Phi_I}(P_n^0)$ using equation 2.13, the empirical mean of $D^*_{\Phi_I}(P_n^0)$ as $P_n D^*_{\Phi_I}(P_n^0) = \frac{1}{n}\sum_{i=1}^n D^*_{\Phi_I}(P_n^0)(B_i)$, and the stopping point $s = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^n D^*_{\Phi_I}(P_n^0)(B_i)^2}}{max(log(n),C)\sqrt{n}}$, where $C$ is a positive constant.
2. Let $m \in \mathbb{Z}$ be the number of iterations, and $P_n^m$ is the updated probability in iteration $m$. Initial $m = 0$. $\delta$ is a positive constant close to zero.

**while** $|P_n D^*_{\Phi_I}(P_n^m)(B_i)| > s$ *and* $|\epsilon| > \delta$ **do**

2.1. calculate the bounds for $\epsilon$ such that the updated probability $\in [0, 1]$.

$$l_\epsilon = max_i[min(-\frac{1}{D^*_{\Phi_I}(P_n^m)(B_i)}, \frac{1 - P_n^m(B_i)}{P_n^m(B_i) D^*_{\Phi_I}(P_n^m)(B_i)}))]$$

$$u_\epsilon = min_i[max(-\frac{1}{D^*_{\Phi_I}(P_n^m)(B_i)}, \frac{1 - P_n^m(B_i)}{P_n^m(B_i) D^*_{\Phi_I}(P_n^m)(B_i)}))]$$

2.2. Construct a least favorable parametric model $\{P_{n,\epsilon}^m : \epsilon\}$ through $P_n^m$ defined as follows:

$$P_{n,\epsilon}^m = C(P_n^m, \epsilon)(1 + \epsilon D^*_{\Phi_I}(P_n^m))P_n^m, \tag{2.16}$$

where

$$C(P, \epsilon) = \frac{1}{\sum_{b\neq 0}(1 + \epsilon D^*_{\Phi_I}(P(b)))P(b)},$$

2.3. calculate

$$\epsilon_n^m = argmax_\epsilon \frac{1}{n}\sum_{b\neq 0} log(P_{n,\epsilon}^m),$$

where $\epsilon \in [l_\epsilon, u_\epsilon]$.
2.4. $m \leftarrow m + 1$, and update $P_n^m \leftarrow P_{n,\epsilon_n^m}^m$.

**end**

3. Denote $P_n^* = P_n^m$ in the final iteration $m$. Calculate TMLE $\Psi_n(P_n^*)$ using equation 2.7, and its efficient influence function $D^*_{\Phi_I}(P_n^*)$ using equation 2.13.
**Result:** TMLE $\Psi_n(P_n^*)$ and efficient influence function $D^*_{\Phi_I}(P_n^*)$.

The proof of the asymptotic efficiency of TMLE is provided in appendix section 2.9.

## 2.5 Simulations

In this section, we show the performance of all the estimators given that their identification assumptions (linear and non-linear constraints) hold true. The estimand $\Psi_f(P)$ and $\Psi_\Phi(P)$ refer to the probability of an individual being captured at least once under the linear constraint $f$ or non-linear constraint $\Phi$. In subsection 2.5, the linear constraint $\Phi_f$ is defined in equation 2.2, the corresponding estimand (equation 2.11)

$$\Psi_f(P) = \frac{f(0)}{f(0) - Pf},$$

and its plug-in estimator

$$\Psi_f(P_n) = \frac{f(0)}{f(0) - \sum_{b\neq 0} f(b)P_n(b)},$$

where $P_n(b)$ is the empirical probability of cell $b$. In subsection 2.5, we provide a summary of all the non-linear constraints defined by equation $\Phi$ and their corresponding estimators. In subsection 2.5, the non-linear constraint $\Phi_I$ is defined by the assumption in equation 2.2. The estimand (equation 2.7)

$$\Psi_I(P) = \frac{1}{1 + \exp\left((-1)^{K+1} \sum_{b\neq 0} f_I(b) \log P(b)\right)}.$$

The plug-in estimator

$$\Psi_I(P_{NP}) = \frac{1}{1 + \exp\left((-1)^{K+1} \sum_{b\neq 0} f_I(b) \log P_n(b)\right)}.$$

In addition, the undersmoothed lasso estimator $\Psi_I(P_{lasso})$ replaces the plugged-in $P_n(b)$ in $\Psi_I(P_{NP})$ with estimated probability in a lasso regression with regularization parameters chosen by the undersmoothing algorithm, the estimator $\Psi_I(P_{lasso\_cv})$ replaces the plugged-in $P_n(b)$ with estimated probability in a lasso regression with regularization parameters optimized by cross-validation. The estimator $\Psi_I(P_{tmle})$ updates the estimated probabilities of $\Psi_I(P_{lasso})$ using targeted learning techniques, and the estimator $\Psi_I(P_{tmle\_cv})$ updates the estimated probabilities of $\Psi_I(P_{lasso\_cv})$ using targeted learning techniques. In addition, we compare the performance of our proposed estimators to existing estimators $\Psi_I(P_{M_0})$ and $\Psi_I(P_{M_t})$.

In subsection 2.5, the non-linear constraint $\Phi_{II}$ is defined by the independence assumption in equation 2.4. The estimand (equation 2.8)

$$\Psi_{II}(P) \equiv \frac{1 - P(B(1) = 0) - P(B(2) = 0) + P(B(1:2) = (0,0))}{1 - P(B(1) = 0) - P(B(2) = 0) + P(B(1) = 0)P(B(2) = 0)},$$

and its plug-in estimator

$$\Psi_{II}(P_n) \equiv \frac{1 - P_n(B(1) = 0) - P_n(B(2) = 0) + P_n(B(1:2) = (0,0))}{1 - P_n(B(1) = 0) - P_n(B(2) = 0) + P_n(B(1) = 0)P_n(B(2) = 0)}.$$

In subsection 2.5, the non-linear constraint $\Phi_{CI}$ is defined by the conditional independence assumption in equation 2.5. The estimand (equation 2.9, $C_0(P)$ is defined in equation 2.10)

$$\Psi_{CI}(P) \equiv \frac{P(B_m = 1, B_j = 1, 0, ..., 0)}{C_0(P)},$$

and its plug-in estimator

$$\Psi_{CI}(P_n)\frac{P_n(B_m = 1, B_j = 1, 0, ..., 0)}{C_0(P_n)}.$$

## Linear identification assumption

The linear identification constraint is $E_{P_{B^*}} f(B^*) = 0$, for any function $f$ such that $f(b = 0) \neq 0$. An example of interest is $f(b) = (-1)^{K+\sum_{k=1}^{K} b_k}$, where K is the total number of samples. In this case, the linear identification assumption is equation 2.2. For example, if $K = 3$, then the constraint (equation 2.2) states:

$$P_{B^*}(b_1, b_2, b_3) = \alpha_0 + \alpha_1 b_1 + \alpha_3 b_2 + \alpha_4 b_1 b_2 + \alpha_5 b_1 b_3 + \alpha_6 b_2 b_3 + \alpha_7 b_1 b_2 b_3$$

and the constraint $E_{P_{B^*}} f(B^*) = 0$ states that $\alpha_7 = 0$. The identified estimator

$$\Psi_f(P_n) = \frac{f(0)}{f(0) - \sum_{b \neq 0} f(b) P_n(b)},$$

where $P_n(b)$ is the observed probability of cell $b$, and the efficient influence curve (equation 2.12) is

$$D_f^* = \frac{f(0)}{(f(0) - \sum_{b \neq 0} f(b) P(b))^2}[f(B) - \sum_{b \neq 0} f(b) P(b)].$$

We use the parameters below to generate the underlying distribution. The assumption that the highest-way interaction term $\alpha_7 = 0$ is satisfied in this setting.

| $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | $\alpha_7$ |
|---|---|---|---|---|---|---|---|
| 0.0725 | 0.03 | 0.01 | 0.04 | 0.01 | 0.02 | 0.02 | 0 |

The simulated true probabilities for all 7 observed cells are:

| P(0,0,1) | P(0,1,0) | P(0,1,1) | P(1,0,0) | P(1,0,1) | P(1,1,0) | P(1,1,1) |
|---|---|---|---|---|---|---|
| 0.1213 | 0.0889 | 0.1429 | 0.1105 | 0.1752 | 0.1428 | 0.2183 |

We can calculate the true value of $\Psi_f(P)$ analytically as $\Psi_f(P_0) = 0.9275$, and draw $10^4$ samples to obtain the asymptotic $\Psi_{f,asym}(P_n) = 0.9316$, asymptotic $\hat{\sigma}^2 = 0.7492$.

Figure 2.1 shows that our estimators perform well when the identification assumption holds true.

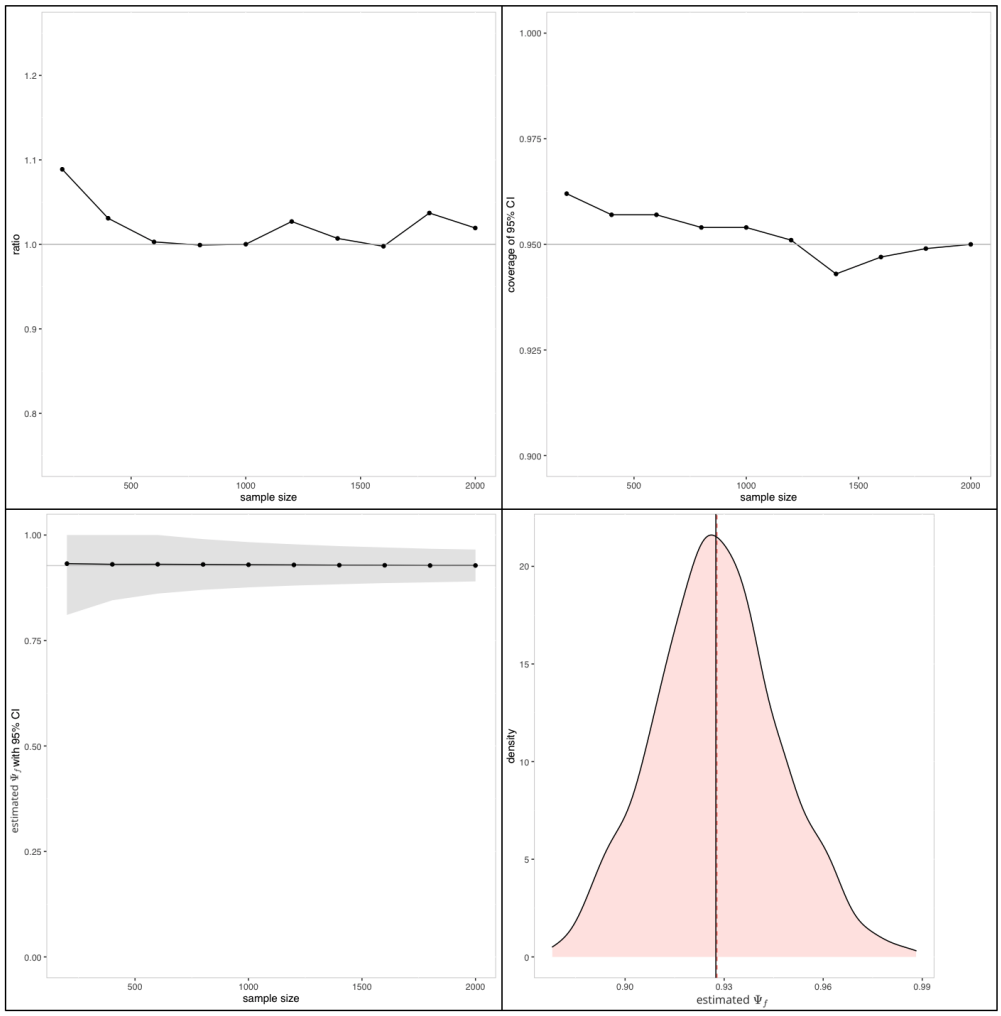


Figure 2.1: Simulation results when the linear identification assumption holds true. **Upper left**: ratio of estimated variance over true variance vs sample size. The line is the ratio at each sample size, and the grey horizontal line is the true value; **Upper right**: coverage of 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve vs sample size. The line is the average coverage; **Lower left**: mean value of estimated $\psi$ and its 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve vs sample size. The line is the mean, and shaded area is the confidence interval and the grey horizontal line is the true value; **Lower right**: Distribution of 1000 estimates of $\psi$ for sample size of 1000, the vertical line is the true value, and the dashed line is the mean value.

## Non-linear identification assumption

We perform simulations for three types of non-linear identification assumptions: 1) K-way interaction term equals zero in log-linear models, 2) independence assumption, and 3) condi-

tional independence assumption. For assumption 1), we provide simulations for the following estimators:

1. Existing estimator defined by model $M_0$, which assumes that in the log-linear model, all the main terms have the same values, and there are no interaction terms [27]: $\Psi_I(P_{M_0})$
2. Existing estimator defined by model $M_t$, which assumes that the log-linear model does not contain interaction terms [11]: $\Psi_I(P_{M_t})$
3. Non-parametric plug-in estimator: $\Psi_I(P_{NP})$
4. Estimator based on undersmoothed lasso regression: $\Psi_I(P_{lasso})$
5. Estimator based on lasso regression with cross-validation: $\Psi_I(P_{lasso\_cv})$
6. TMLE based on $\Psi_I(P_{lasso})$: $\Psi_I(P_{tmle})$
7. TMLE based on $\Psi_I(P_{lasso\_cv})$: $\Psi_I(P_{tmle\_cv})$

Model $M_0$ and $M_t$ were calculated using R package "RCapture" (version 1.4-3) [28].

## Non-linear identification assumption: k-way interaction term equals zero

Given that we have three samples $B_1, B_2$ and $B_3$, the identification assumption is that in the log-linear model

$$log(E_{B^*}(b)) = \alpha_0 + \alpha_1 b_1 + \alpha_2 b_2 + \alpha_3 b_3 + \alpha_4 b_1 b_2 + \alpha_5 b_1 b_3 + \alpha_6 b_2 b_3 + \alpha_7 b_1 b_2 b_3,$$

$\alpha_7 = 0$ (equation 2.2). Here $E_{B^*}(b)$ is the count of observations in cell $b = (b_1, b_2, b_3)$, and $b_i = 1$ if the subject is captured by sample $i, i = 1, 2, 3$. The parameter $\Psi_I(P)$ is identified in equation 2.7, and its influence curve $D^*_{\Phi_I}(P)$ is identified in equation 2.13.

In this section, we evaluated both our proposed estimators as well as two existing parametric estimators. The first estimator we evaluated is the $M_t$ model [29], where $P_{ij}$ denotes the probability that subject $i$ is captured by sample $j$, and modeled as $log(P_{ij}) = \mu_j$ for subject $i$ and sample $j$. This estimator assumes that all the interaction terms are zero in the log-linear model, and only use the main term variables in model training. The second estimator is $M_0$ [27], with the formula $log(P_{ij}) = \alpha$, where $\alpha$ is a constant.

### Log-linear model with main term effects

We used the parameters below to generate the underlying distribution. The assumption that the k-way interaction term $\alpha_7 = 0$ is satisfied in this setting. The model assumptions for $M_0$ and $M_t$ are also satisfied in this setting.

| $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | $\alpha_7$ |
|---|---|---|---|---|---|---|---|
| -0.9398 | -1 | -1 | -1 | 0 | 0 | 0 | 0 |

The simulated observed probabilities for all 7 observed cells are:

| P(0,0,1) | P(0,1,0) | P(0,1,1) | P(1,0,0) | P(1,0,1) | P(1,1,0) | P(1,1,1) |
|---|---|---|---|---|---|---|
| 0.2359 | 0.2359 | 0.0868 | 0.2359 | 0.0868 | 0.0868 | 0.0319 |

Under this setting, the true value of $\Psi_I(P)$ can be calculated analytically as $\Psi_I(P_0) = 0.6093$. Figure 2.2 and table 2.1 shows the performance of estimators $\Psi_I(P_{M_0})$, $\Psi_I(P_{M_t})$, $\Psi_I(P_{NP})$, $\Psi_I(P_{lasso})$, $\Psi_I(P_{lasso\_cv})$, $\Psi_I(P_{tmle})$, and $\Psi_I(P_{tmle\_cv})$. Table 2.2 shows that the model fit statistics for$\Psi_I(P_{M_0})$ and $\Psi_I(P_{M_t})$.

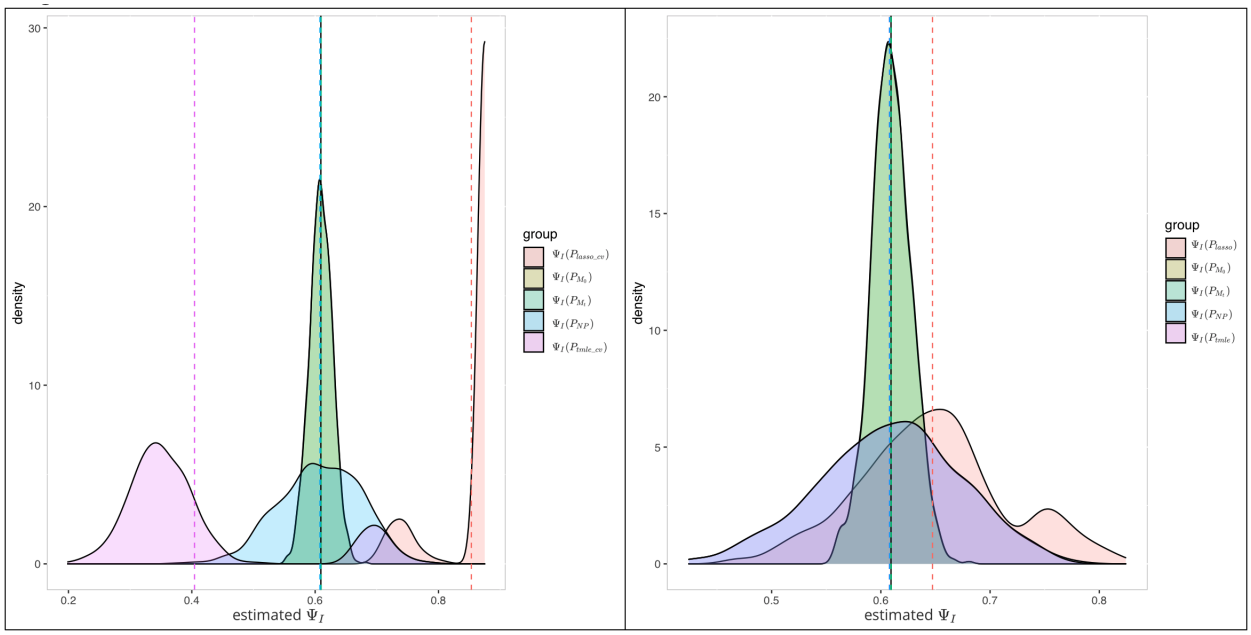


Figure 2.2: **Left**: Distribution of 1000 estimates of $\psi$ for sample size of 1000 with estimators $\Psi_I(P_{lasso\_cv})$ and $\Psi_I(P_{tmle\_cv})$, the black vertical line is the true value, and the dashed vertical lines represent mean values for each estimator;
**Right**: Distribution of 1000 estimates of $\psi$ for sample size of 1000 with estimators $\Psi_I(P_{lasso})$ and $\Psi_I(P_{tmle})$, the black vertical line is the true value, and the dashed vertical lines represent mean values for each estimator.

| Estimator | Average $\psi$ | Lower 95% CI | Upper 95% CI | Coverage(%) |
|---|---|---|---|---|
| $\Psi_I(P_{NP})$ | 0.6078 | 0.4791 | 0.7366 | 95 |
| $\Psi_I(P_{lasso})$ | 0.6473 | 0.5237 | 0.7708 | 84.2 |
| $\Psi_I(P_{tmle})$ | 0.6078 | 0.4806 | 0.735 | 93.7 |
| $\Psi_I(P_{lasso\_cv})$ | 0.8534 | 0.7984 | 0.9083 | 4.4 |
| $\Psi_I(P_{tmle\_cv})$ | 0.4044 | 0.2988 | 0.5101 | 13.6 |
| $\Psi_I(P_{M_0})$ | 0.6093 | 0.5659 | 0.6521 | 97.8 |
| $\Psi_I(P_{M_t})$ | 0.6096 | 0.5662 | 0.6525 | 97.8 |

Table 2.1: Estimated $\hat{\psi}$ and 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve and its coverage by each estimator. $\Psi_I(P_{NP})$ is the plug-in maximum likelihood estimator, $\Psi_I(P_{lasso})$ uses probabilities estimated from undersmoothed lasso regression, $\Psi_I(P_{tmle})$ is the TMLE based on $\Psi_I(P_{lasso})$, $\Psi_I(P_{lasso\_cv})$ uses probabilities estimated from lasso regression with regularization term optimized by cross-validation, $\Psi_I(P_{tmle\_cv})$ is the TMLE based on $\Psi_I(P_{lasso\_cv})$, $\Psi_I(P_{M_0})$ and $\Psi_I(P_{M_t})$ are existing estimators defined in section 2.5. True $\psi_0 = 0.6093$.

| Estimator | Df | AIC | BIC |
|---|---|---|---|
| $\Psi_I(P_{M_0})$ | 5 | 56.532 | 66.348 |
| $\Psi_I(P_{M_t})$ | 3 | 58.034 | 77.665 |

Table 2.2: model fit statistics for $\Psi_I(P_{M_0})$ and $\Psi_I(P_{M_t})$ estimators.

Figure 2.2 and table 2.1 compared the performance of two base learners: lasso regression with regularization term chosen by cross-validation ($\Psi_I(P_{lasso\_cv})$), and undersmoothed lasso regression ($\Psi_I(P_{lasso})$).The $\Psi_I(P_{lasso\_cv})$ estimator is significantly biased from the true value of $\psi$, and the TMLE estimator based on it, $\Psi_I(P_{tmle\_cv})$ estimator is also biased. And although the $\Psi_I(P_{lasso})$ estimator is biased, the TMLE based on it, $\Psi_I(P_{tmle})$ estimator ($mean = 0.6078$, $coverage = 93.7\%$) is able to adjust the bias and give a fit as good as the NPMLE estimator ($mean = 0.6078$, $coverage = 95\%$). Thus, in the section below we will only use the undersmooting lasso estimator $\Psi_I(P_{lasso})$ as the base learner.

**Log-linear model with main term and interaction term effects**

We used the parameters below to generate the underlying distribution. The assumption that the k-way interaction term $\alpha_7 = 0$ is satisfied in this setting.

| $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | $\alpha_7$ |
|---|---|---|---|---|---|---|---|
| -0.9194 | -1 | -1 | -1 | -0.1 | -0.1 | -0.1 | 0 |

The simulated observed probabilities for all 7 observed cells are:

| P(0,0,1) | P(0,1,0) | P(0,1,1) | P(1,0,0) | P(1,0,1) | P(1,1,0) | P(1,1,1) |
|---|---|---|---|---|---|---|
| 0.2440 | 0.2440 | 0.0812 | 0.2440 | 0.0812 | 0.0812 | 0.0245 |

Under this setting, the true value of $\Psi_I(P)$ can be calculated analytically as $\Psi_I(P_0) = 0.6013$. The model assumptions of $\Psi_I(P_{M_0})$ and $\Psi_I(P_{M_t})$ are violated, and their estimates are biased (estimated mean $\hat{\Psi}_I(P_{M_0}) = \hat{\Psi}_I(P_{M_t}) = 0.572, bias = \Psi_I(P_0) - \hat{\Psi}_I(P_{M_t}) = 0.6013 - 0.5720 = 0.0293$). The coverage of their 95% confidence interval is also low (81.5%). Table 2.4 shows that the model fit statistics for$\Psi_I(P_{M_0})$ and $\Psi_I(P_{M_t})$. As the $\Psi_I(P_{NP})$ and $\Psi_I(P_{tmle})$ estimators do not have model assumptions, they are unbiased and the coverage of their 95% asymptotic confidence intervals are close to 95%. (94.7% for $\Psi_I(P_{NP})$ and 93.7% for $\Psi_I(P_{tmle})$), shown in figure 2.3 and table 2.3.

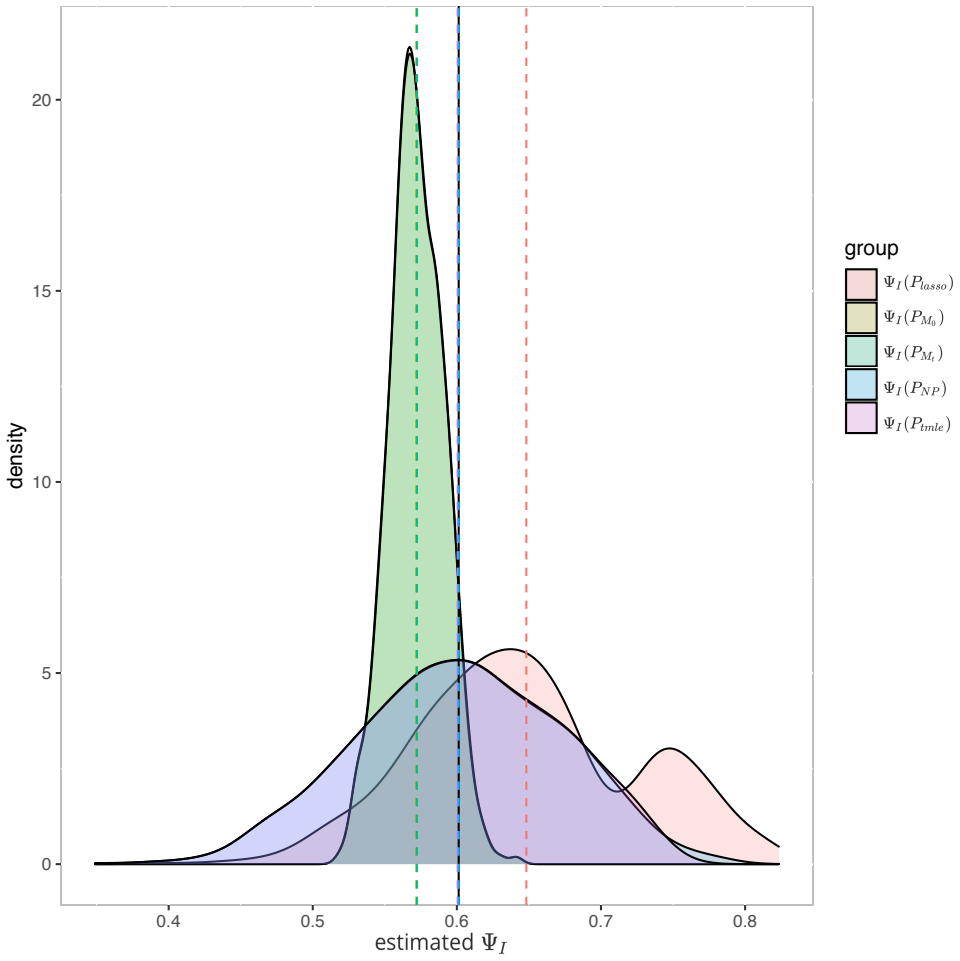


Figure 2.3: Distribution of 1000 estimates of $\psi$ for sample size of 1000, the black vertical line is the true value, and the dashed vertical lines represent mean values for each estimator..

| Estimator | Average $\psi$ | Lower 95% CI | Upper 95% CI | Coverage(%) |
|---|---|---|---|---|
| $\Psi_I(P_{NP})$ | 0.6013 | 0.4625 | 0.7401 | 94.7 |
| $\Psi_I(P_{lasso})$ | 0.6482 | 0.5159 | 0.7805 | 78.4 |
| $\Psi_I(P_{tmle})$ | 0.6005 | 0.4653 | 0.7358 | 93.7 |
| $\Psi_I(P_{M_0})$ | 0.572 | 0.5274 | 0.6163 | 81.5 |
| $\Psi_I(P_{M_t})$ | 0.5722 | 0.5277 | 0.6166 | 81.5 |

Table 2.3: Estimated $\hat{\psi}$ and 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve and its coverage by each estimator. $\Psi_I(P_{NP})$ is the plug-in maximum likelihood estimator, $\Psi_I(P_{lasso})$ uses probabilities estimated from undersmoothed lasso regression, $\Psi_I(P_{tmle})$ is the TMLE based on $\Psi_I(P_{lasso})$, $\Psi_I(P_{M_0})$ and $\Psi_I(P_{M_t})$ are existing estimators defined in section 2.5. True $\Psi_I(P_0) = 0.6013$.

| Estimator | Df | AIC | BIC |
|---|---|---|---|
| $\Psi_I(P_{M_0})$ | 5 | 55.499 | 65.314 |
| $\Psi_I(P_{M_t})$ | 3 | 58.440 | 78.071 |

Table 2.4: model fit statistics for$\Psi_I(P_{M_0})$ and $\Psi_I(P_{M_t})$ estimators.

**Log-linear model with empty cells**

When the probability for some cells are close to zero, in the finite sample case, there are likely to be empty cells in the observed data. For example, if the probability of a subject being caught in all 3 samples, represented as $P(1,1,1)$, is less than $10^{-6}$ and the total number of unique subjects caught by any of the 3 samples is less than $10^3$, then it's likely that we will not observe any subject being caught three times, i.e., cell $P(1,1,1)$ will likely be empty.

We used the parameters below to generate the underlying distribution. The assumption that the k-way interaction term $\alpha_7 = 0$ is satisfied in this setting.

| $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | $\alpha_7$ |
|---|---|---|---|---|---|---|---|
| -0.4578 | -1 | -2 | -3 | -1 | -1 | -1 | 0 |

The simulated observed probabilities for all 7 observed cells are:

| P(0,0,1) | P(0,1,0) | P(0,1,1) | P(1,0,0) | P(1,0,1) | P(1,1,0) | P(1,1,1) |
|---|---|---|---|---|---|---|
| 0.0857 | 0.2331 | 0.0043 | 0.6336 | 0.0116 | 0.0315 | 2e-04 |

Figure 2.4 and table 2.5 show the performance of the estimators when there's no observation in cell $(1,1,1)$, .

Figure 2.4 shows that all the existing estimators are biased when empty cells of $(1,1,1)$ exist. Table 2.5 shows that the coverage of the 95% asymptotic confidence intervals for $\Psi_I(P_{tmle})$ is the highest (58.8%), and the coverage of the 95% asymptotic confidence intervals for $\Psi_I(P_{M_0})$, $\Psi_I(P_{M_t})$ are both 0, due to the estimation bias and narrow range of intervals. Table 2.6 shows the model fit statistics for$\Psi_I(P_{M_0})$ and $\Psi_I(P_{M_t})$.

| Estimator | Average $\psi$ | Lower 95% CI | Upper 95% CI | Coverage(%) |
|---|---|---|---|---|
| $\Psi_I(P_{lasso})$ | 0.5565 | 0.1956 | 0.9174 | 20 |
| $\Psi_I(P_{tmle})$ | 0.2308 | 0.0875 | 0.374 | 58.8 |
| $\Psi_I(P_{M_0})$ | 0.1318 | 0.0995 | 0.169 | 0 |
| $\Psi_I(P_{M_t})$ | 0.1729 | 0.1313 | 0.2201 | 0 |

Table 2.5: Estimated $\hat{\psi}$, 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve and its coverage of the estimators in figure 2.4. $\Psi_I(P_{lasso})$ uses probabilities estimated from undersmoothed lasso regression, $\Psi_I(P_{tmle})$ is the TMLE based on $\Psi_I(P_{lasso})$, $\Psi_I(P_{M_0})$ and $\Psi_I(P_{M_t})$ are existing estimators defined in section 2.5.True $\Psi_I(P_0) = 0.3674$.

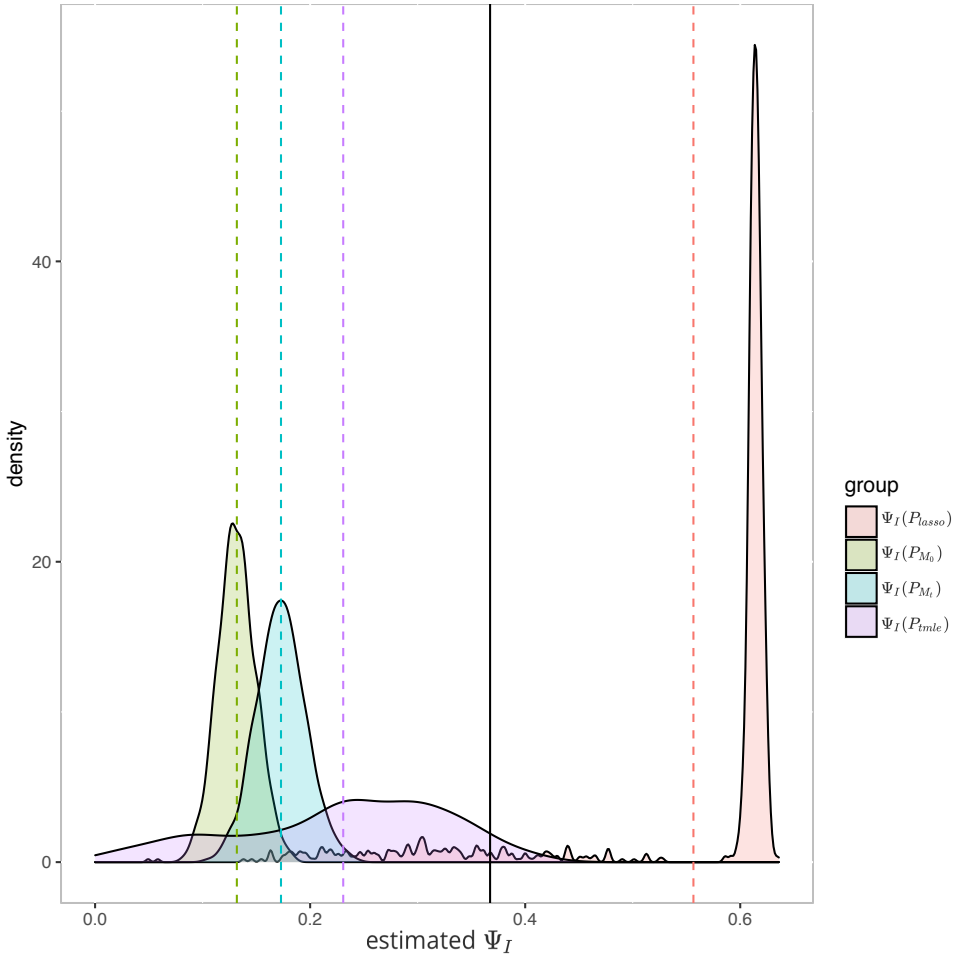


Figure 2.4: Distribution of 1000 estimates of $\psi$ for sample size of 1000, the black vertical line is the true value of $\Psi_I(P_0) = 0.3674$, and the dashed vertical lines represent mean values for each estimator.

## Non-linear identification assumption: independence

Given 3 samples, we assume that the first and second sample $B_1, B_2$ are independent of each other, that is, $P^*(B_1 = 1, B_2 = 1) = P^*(B_1 = 1) \times P^*(B_2 = 1)$. Then we have $\Psi_{II}(P)$

| Estimator | Df | AIC | BIC |
|---|---|---|---|
| $\Psi_I(P_{M_0})$ | 5 | 501.562 | 511.378 |
| $\Psi_I(P_{M_t})$ | 3 | 45.506 | 65.137 |

Table 2.6: model fit statistics for$\Psi_I(P_{M_0})$ and $\Psi_I(P_{M_t})$ estimators.

identified in equation 2.8, and the influence curve $D^*_{II}(P)$ identified in equation 2.14.

Here we illustrate the performance of our estimators with the following underlying distribution:

$$
\begin{aligned}
&P(B_1 = 1) = 0.1 \\
&P(B_2 = 1|B_1 = 1) = 0.2 \\
&P(B_2 = 1|B_1 = 0) = 0.2 \\
&P(B_3 = 1|B_1 = 1) = 0.25 \\
&P(B_3 = 1|B_1 = 0) = 0.3
\end{aligned}
$$

The probability distribution for all 7 observed cells are:

| P(0,0,1) | P(0,1,0) | P(0,1,1) | P(1,0,0) | P(1,0,1) | P(1,1,0) | P(1,1,1) |
|---|---|---|---|---|---|---|
| 0.4355 | 0.2540 | 0.1089 | 0.1210 | 0.0403 | 0.0302 | 0.0101 |

We can calculate the true value of $\Psi_{II}(P)$ analytically as $\Psi_{II}(P_0) = 0.4960$, and draw $10^6$ samples to obtain the asymptotic $\Psi_{II}(P_n) = 0.4963$, asymptotic $\sigma^2 = \frac{1}{n}\sum D^{*2} = 4.2657$.

Figure 2.5 shows that our estimators perform well when the independence assumptions hold true.

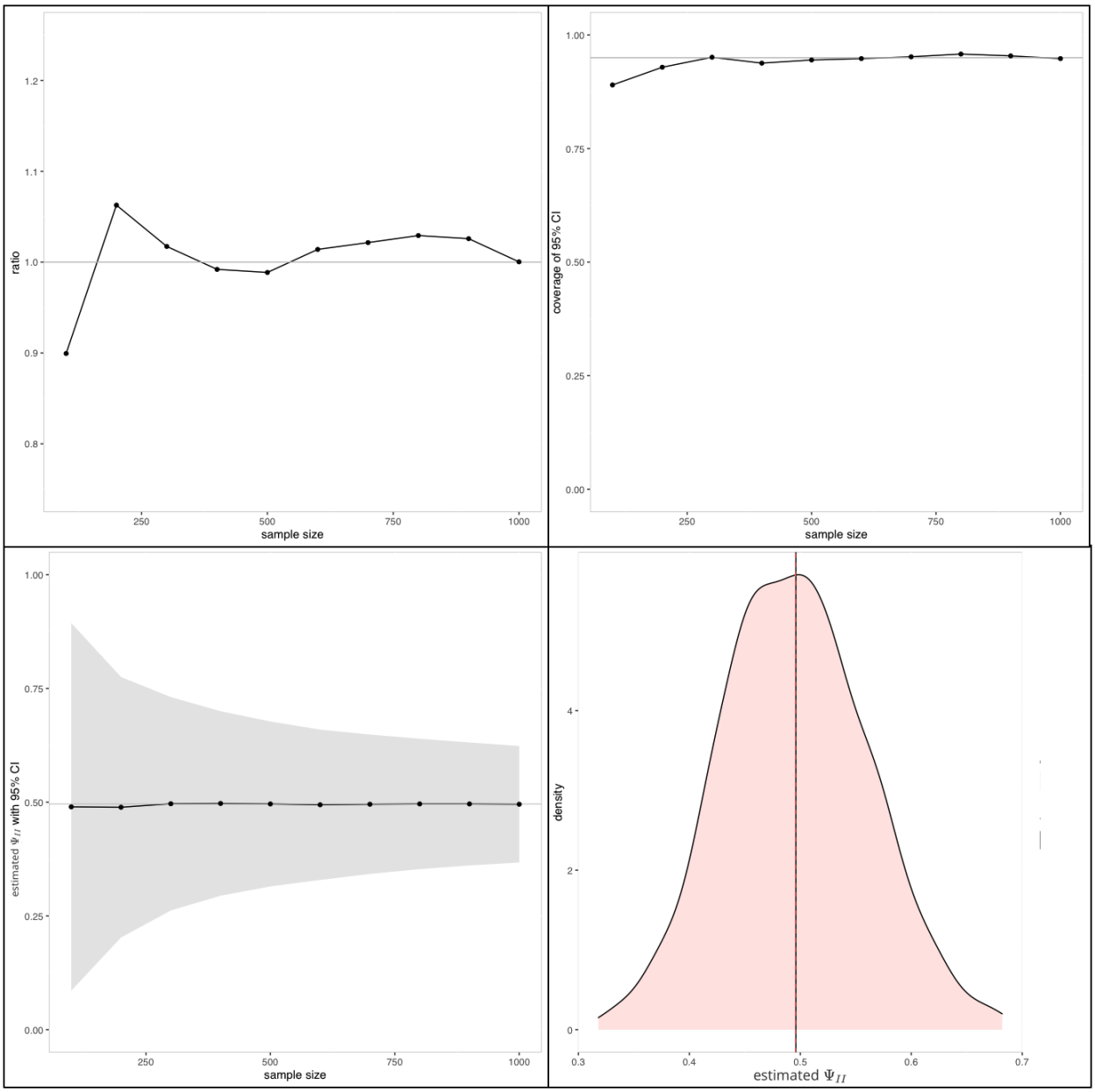


Figure 2.5: Simulation results when the independence identification assumption holds true. **Upper left**: ratio of estimated variance over true variance vs sample size. The line is the ratio at each sample size, and the grey horizontal line is the true value; **Upper right**: coverage of 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve vs sample size. The line is the average coverage; **Lower left**: mean value of estimated $\psi$ and its 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve vs sample size. The line is the mean, and shaded area is the confidence interval and the grey horizontal line is the true value; **Lower right**: Distribution of 1000 estimates of $\psi$ for sample size of 1000, the vertical line is the true value, and the dashed line is the mean value.

## Non-linear identification assumption: conditional independence

Given 3 samples, we assume that the third sample $B_3$ is conditionally independent on the second sample $B_2$, that is, $P^*(B_3 = 1|B_2 = b_2, B_1 = 0) = P^*(B_3 = 1|B_1 = 0)$. Then we can derive that $P^*(0,0,0) = \frac{P^*(0,1,0)P^*(0,0,1)}{P^*(0,1,1)}$ and $\Psi_{CI}(P) = \frac{P(0,1,1)}{P(0,1,1)+P(0,1,0)P(0,0,1)}$ from equation 2.9, and the efficient influence curve can be derived from equation 2.15 as

$$
\begin{aligned}
&D^*_{\Phi_{CI}}(P) = \\
&\quad \frac{P(0,1,0)P(0,0,1)}{[P(0,1,1)+P(0,1,0)P(0,0,1)]^2}[I(0,1,1) - P(0,1,1)] - \\
&\quad \frac{P(0,0,1)P(0,1,1)}{[P(0,1,1)+P(0,1,0)P(0,0,1)]^2}[I(0,1,0) - P(0,1,0)] - \\
&\quad \frac{P(0,1,0)P(0,1,1)}{[P(0,1,1)+P(0,1,0)P(0,0,1)]^2}[I(0,0,1) - P(0,0,1)]
\end{aligned}
$$

Here we illustrate the performance of our estimators with the following underlying distribution:

$$
\begin{aligned}
&P(B_1 = 1) = 0.1 \\
&P(B_2 = 1|B_1 = 1) = 0.2 \\
&P(B_2 = 1|B_1 = 0) = 0.15 \\
&P(B_3 = 1|B_1 = 1) = P(B_3 = 1|B_2 = 1, B_1 = 1) = 0.25 \\
&P(B_3 = 1|B_1 = 1) = P(B_3 = 1|B_2 = 0, B_1 = 1) = 0.25 \\
&P(B_3 = 1|B_1 = 0) = P(B_3 = 1|B_2 = 1, B_1 = 0) = 0.2 \\
&P(B_3 = 1|B_1 = 0) = P(B_3 = 1|B_2 = 0, B_1 = 0) = 0.2
\end{aligned}
$$

The probability distribution for all 7 observed cells $P(B_1, B_2, B_3)$ are:

| P(0,0,1) | P(0,1,0) | P(0,1,1) | P(1,0,0) | P(1,0,1) | P(1,1,0) | P(1,1,1) |
|---|---|---|---|---|---|---|
| 0.3943 | 0.2784 | 0.0696 | 0.1546 | 0.0515 | 0.0387 | 0.0129 |

We can derive the true $\Psi_{CI}(P)$ analytically as $\Psi_{CI}(P_0) = 0.388$, and draw $10^6$ samples to obtain the asymptotic $\hat{\Psi}_{CI}(P_n) = 0.3887$, asymptotic $\hat{\sigma}^2 = 1.0958$.

Figure 2.6 shows that our estimator performs well when the conditional independence assumptions are met.

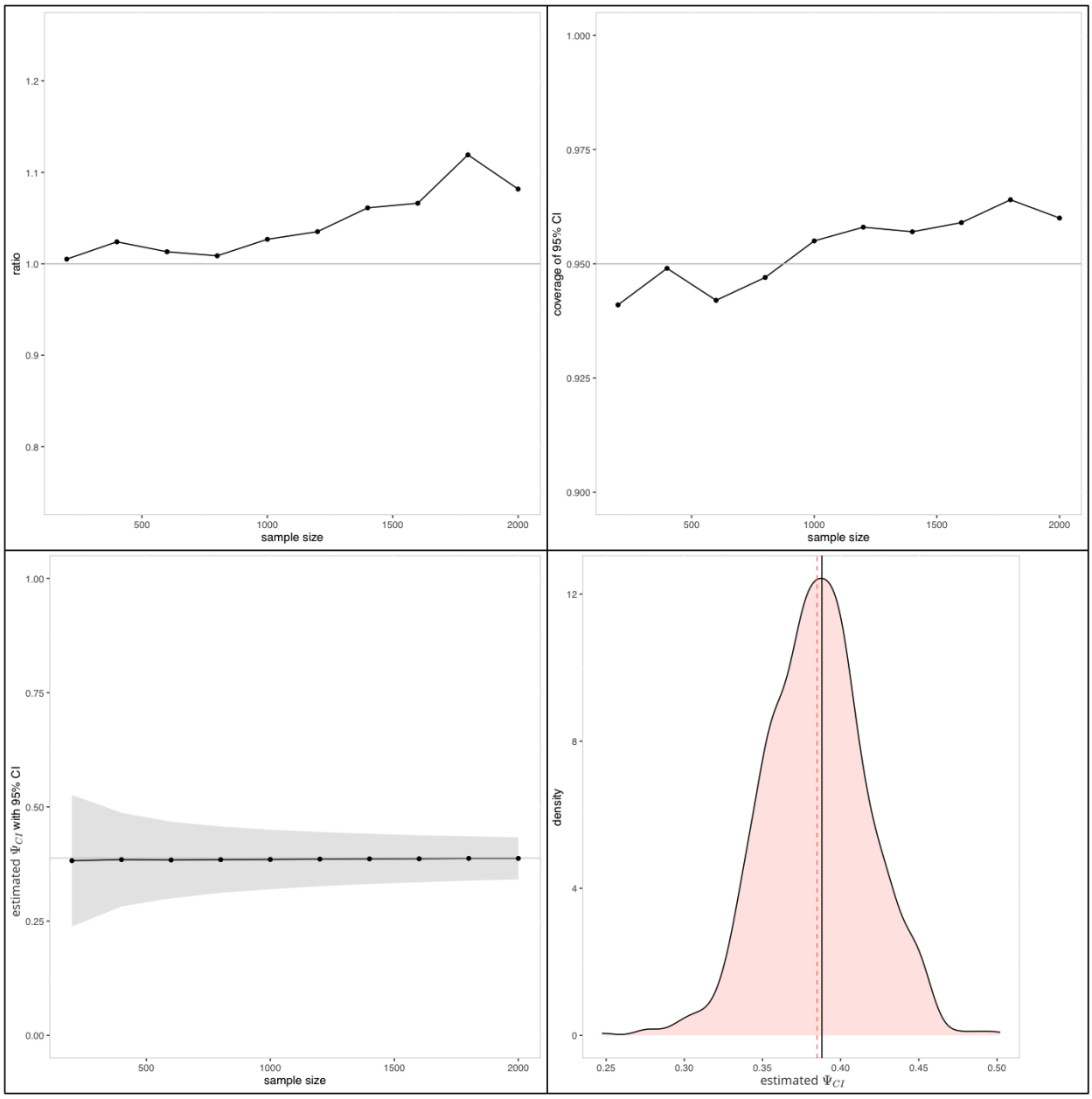


Figure 2.6: Simulation results when the conditional independence identification assumption holds true. **Upper left**: ratio of estimated variance over true variance vs sample size. The line is the ratio at each sample size, and the grey horizontal line is the true value; **Upper right**: coverage of 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve vs sample size. The line is the average coverage; **Lower left**: mean value of estimated $\psi$ and its 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve vs sample size. The line is the mean, and shaded area is the confidence interval and the grey horizontal line is the true value; **Lower right**: Distribution of 1000 estimates of $\psi$ for sample size of 1000, the vertical line is the true value, and the dashed line is the mean value.

## 2.6 Evaluating the sensitivity of identification bias to violations of the assumed constraint

In this section, we use simulations to show the sensitivity of identification bias for all the estimators with identification assumptions violations.

### Violation of linear assumptions

In this section we analyzed the same problem stated in section 2.5, but the identification assumption is violated.

We used the parameters below to generate the underlying distribution. The assumption that the k-way interaction term $\alpha_7 = 0$ is violated in this setting($\alpha_7 = 0.2$).

| $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | $\alpha_7$ |
|---|---|---|---|---|---|---|---|
| 0.11 | 0.1 | 0.05 | 0.08 | -0.2 | -0.2 | -0.1 | 0.2 |

The simulated observed probabilities for all 7 cells are:

| P(0,0,1) | P(0,1,0) | P(0,1,1) | P(1,0,0) | P(1,0,1) | P(1,1,0) | P(1,1,1) |
|---|---|---|---|---|---|---|
| 0.2135 | 0.1798 | 0.0449 | 0.2360 | 0.1011 | 0.1978 | 0.0449 |

We can calculate the true value of $\Psi$ analytically as $\Psi_0 = 0.8900$, and draw $10^5$ samples to obtain the asymptotic $\Psi_f(P_n) = 0.7419$, asymptotic $\hat{\sigma}^2 = 0.2662$. The asymptotic value is biased by 0.1481.

Figure 2.7 shows that the linear estimator is biased when the identification assumption is violated.

## Violation of non-linear assumption: independence

Given 3 samples, we suppose that the first and second sample $B_1, B_2$ are independent of each other, that is, $P^*(B_1 = 1, B_2 = 1) = P^*(B_1 = 1) \times P^*(B_2 = 1)$. Then we can compute $\hat{\psi}$ and its efficient influence curve as stated above.

Here we illustrate the performance of our estimators with the following underlying distribution:

$$
\begin{aligned}
&P(B_1 = 1) = 0.5 \\
&P(B_2 = 1|B_1 = 1) = 0.6 \\
&P(B_2 = 1|B_1 = 0) = 0.5 \\
&P(B_3 = 1) = 0.5
\end{aligned}
$$

The probability for all 7 observed cells is given by:

| P(0,0,1) | P(0,1,0) | P(0,1,1) | P(1,0,0) | P(1,0,1) | P(1,1,0) | P(1,1,1) |
|---|---|---|---|---|---|---|
| 0.1429 | 0.1429 | 0.1429 | 0.1143 | 0.1143 | 0.1714 | 0.1714 |

We can calculate the true value of $\Psi$ analytically as $\Psi_0 = 0.875$, and draw $10^6$ samples to obtain the asymptotic $\Psi_{II}(P_n) = 0.9544$, asymptotic $\sigma^2 = 0.4415$, the asymptotic value is biased by 0.0794.

Figure 2.8 shows that violation of the independence assumption will create a significant bias in the results.

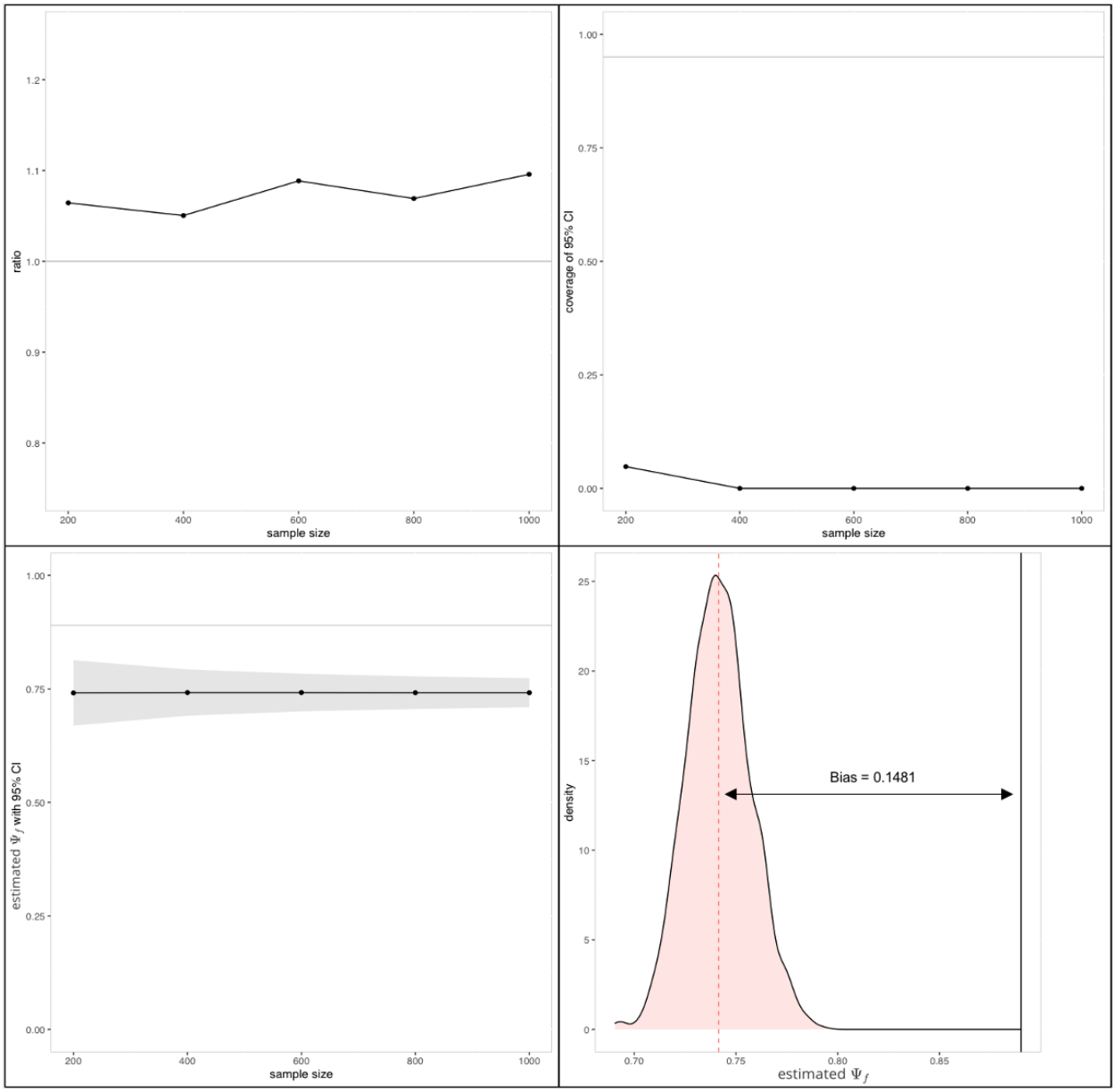


Figure 2.7: Simulation results when the linear identification assumption is violated. **Upper left**: ratio of estimated variance over true variance vs sample size. The line is the ratio at each sample size, and the grey horizontal line is the true value; **Upper right**: coverage of 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve vs sample size. The line is the average coverage; **Lower left**: mean value of estimated $\psi$ and its 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve vs sample size. The line is the mean, and shaded area is the confidence interval and the grey horizontal line is the true value; **Lower right**: Distribution of 1000 estimates of $\psi$ for sample size of 1000, the vertical line is the true value, and the dashed line is the mean value.

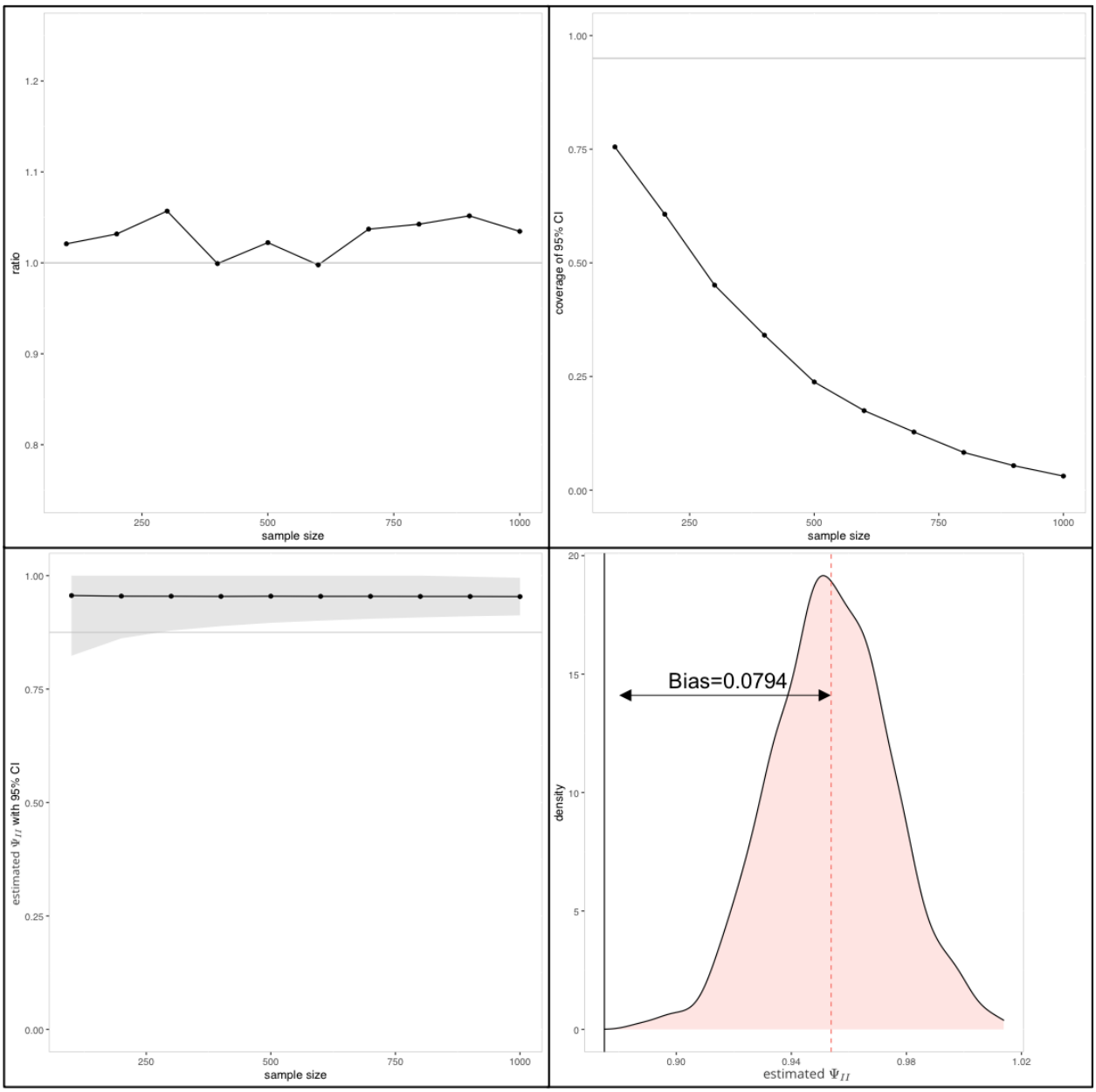


Figure 2.8: Simulation results when the independence identification assumption is violated. **Upper left**: ratio of estimated variance over true variance vs sample size. The line is the ratio at each sample size, and the grey horizontal line is the true value; **Upper right**: coverage of 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve vs sample size. The line is the average coverage; **Lower left**: mean value of estimated $\psi$ and its 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve vs sample size. The line is the mean, and shaded area is the confidence interval and the grey horizontal line is the true value; **Lower right**: Distribution of 1000 estimates of $\psi$ for sample size of 1000, the vertical line is the true value, and the dashed line is the mean value.

## Violation of non-linear assumption: conditional independence

Same as above, our assumption is that given 3 samples, the third sample $B_3$ is independent on the second sample $B_2$ conditional on the first sample $B_1$. Then we can derive the influence curve same as above.

Here we illustrate the performance of our estimators when the identification assumption does not hold.

The simulated distribution is as follows:

$$
\begin{aligned}
&P(B_1 = 1) = 0.1 \\
&P(B_2 = 1|B_1 = 1) = 0.2 \\
&P(B_2 = 1|B_1 = 0) = 0.15 \\
&P(B_3 = 1|B_1 = 1, B_2 = 1) = 0.10 \\
&P(B_3 = 1|B_1 = 1, B_2 = 0) = 0.50 \\
&P(B_3 = 1|B_1 = 0, B_2 = 1) = 0.20 \\
&P(B_3 = 1|B_1 = 0, B_2 = 0) = 0.50
\end{aligned}
$$

The probability distribution for all 7 observed cells are:

| P(0,0,1) | P(0,1,0) | P(0,1,1) | P(1,0,0) | P(1,0,1) | P(1,1,0) | P(1,1,1) |
|---|---|---|---|---|---|---|
| 0.6194 | 0.1749 | 0.0437 | 0.0648 | 0.0648 | 0.0291 | 0.0032 |

The true value of $\Psi$ is $\Psi_0 = 0.6175$, and we drew $10^6$ samples and obtained the asymptotic $\Psi_{CI}(P_n) = 0.2931$, asymptotic $\sigma^2 = 1.2349$, the asymptotic value is biased by 0.3244.

Figure 2.9 shows that with a strong violation of conditional independence assumptions, the estimated values will have significant biases.

## Violation of non-linear assumption: K-way interaction term equals zero

Here we illustrate the performance of our estimators when the identification assumption that the k-way interaction term $\alpha_7 = 0$ does not hold. We used the parameters below to generate the underlying distribution. The underlying distribution follows the same model as in section 2.5.

| $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | $\alpha_7$ |
|---|---|---|---|---|---|---|---|
| -1.6333 | 0 | 0 | 0 | -1 | -2 | -0.5 | 1 |

The simulated observed probabilities for all 7 observed cells are:

| P(0,0,1) | P(0,1,0) | P(0,1,1) | P(1,0,0) | P(1,0,1) | P(1,1,0) | P(1,1,1) |
|---|---|---|---|---|---|---|
| 0.2386 | 0.2386 | 0.0878 | 0.2386 | 0.0323 | 0.1447 | 0.0196 |

Figure 2.10 and table 2.7 show that when the assumption is heavily violated, all estimators are significantly biased. The true value of $\psi = 0.8074$. The bias for $\Psi_I(P_{NP})$ estimator is 0.2032, for $\Psi_I(P_{lasso})$ estimator is 0.1372, for $\Psi_I(P_{tmle})$ is 0.2055, for $\Psi_I(P_{M_0})$ estimator is 0.2217, and for $\Psi_I(P_{M_t})$ estimator is 0.2185. The coverage of 95% asymptotic confidence intervals for $\Psi_I(P_{NP})$ estimator is 26%, for $\Psi_I(P_{lasso})$ estimator is 59%, for $\Psi_I(P_{tmle})$ is 22%, for $\Psi_I(P_{M_0})$ and $\Psi_I(P_{M_t})$ is 0%. Table 2.10 shows the information on $M_0, M_t$ models.

| Estimator | average $\psi$ | lower 95% CI | upper 95% CI | coverage(%) |
|---|---|---|---|---|
| $\Psi_I(P_{NP})$ | 0.6042 | 0.4491 | 0.7592 | 26.3 |
| $\Psi_I(P_{lasso})$ | 0.6702 | 0.5261 | 0.8142 | 58.6 |
| $\Psi_I(P_{tmle})$ | 0.6019 | 0.4516 | 0.7523 | 21.9 |
| $\Psi_I(P_{M_0})$ | 0.5857 | 0.5416 | 0.6295 | 0 |
| $\Psi_I(P_{M_t})$ | 0.5889 | 0.5447 | 0.6328 | 0 |

Table 2.7: Estimated $\hat{\psi}$ and 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve and its coverage by each estimator. True $\Psi_0 = 0.8074$.

| Estimator | df | AIC | BIC |
|---|---|---|---|
| $\Psi_I(P_{M_0})$ | 5 | 157.720 | 167.535 |
| $\Psi_I(P_{M_t})$ | 3 | 128.623 | 148.254 |

Table 2.8: model fit statistics for$\Psi_I(P_{M_0})$ and $\Psi_I(P_{M_t})$ estimators.

## Violation of all identification assumptions

In the sections above we showed that if an identification assumption is violated, its corresponding estimators will be biased and their asymptotic confidence intervals will have low coverage. In this section, all the identification assumptions are violated in the simulated distribution, and we analyzed the performance of all the estimators discussed above in this setting.

The underlying distribution of three samples $B_1, B_2, B_3$ follows the same log-linear model in section 2.5, with the parameters as below.

| $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | $\alpha_7$ |
|---|---|---|---|---|---|---|---|
| -1.1835 | -1 | -1 | -1 | -1.5 | -1 | 2 | 1 |

The simulated observed probability for all 7 observed cells is given by:

| P(0,0,1) | P(0,1,0) | P(0,1,1) | P(1,0,0) | P(1,0,1) | P(1,1,0) | P(1,1,1) |
|---|---|---|---|---|---|---|
| 0.1624 | 0.1624 | 0.0133 | 0.1624 | 0.0220 | 0.4414 | 0.0362 |

Figure 2.11 and table 2.9 show that all estimators are significantly biased, when none of their identification assumption holds. The true value of $\psi = 0.6938$. Under the assumption that sample $B_1$ and $B_2$ are independent conditional on $B_3$ ("Conditional Independence" in table 2.9), the plug-in estimator has a bias of -0.2497. Under the assumption that sample $B_1$ and $B_2$ are independent ("Independence" in table 2.9), the plug-in estimator has a bias of 0.3760. Under the Under the assumption that the 3-way additive interaction term in the linear model equals zero ("K-way additive interaction equals zero" in table 2.9), the plug-in estimator has a bias of -0.2627. And under the assumption that the 3-way multiplicative interaction term in the log-linear model equals zero ("K-way multiplicative interaction equals zero" in table 2.9), the $\Psi_I(P_{NP})$ estimator has a bias of 0.2517, the $\Psi_I(P_{lasso})$ estimator has a bias of -0.1573, the $\Psi_I(P_{tmle})$ has a bias of -0.1867, the $\Psi_I(P_{M_0})$ estimator has a bias of -0.1017, and the $\Psi_I(P_{M_t})$ estimator has a bias of -0.1446. The coverage of 95% asymptotic confidence intervals for all the estimators are far below 95%. Table 2.10 shows the information on $M_0, M_t$ models.

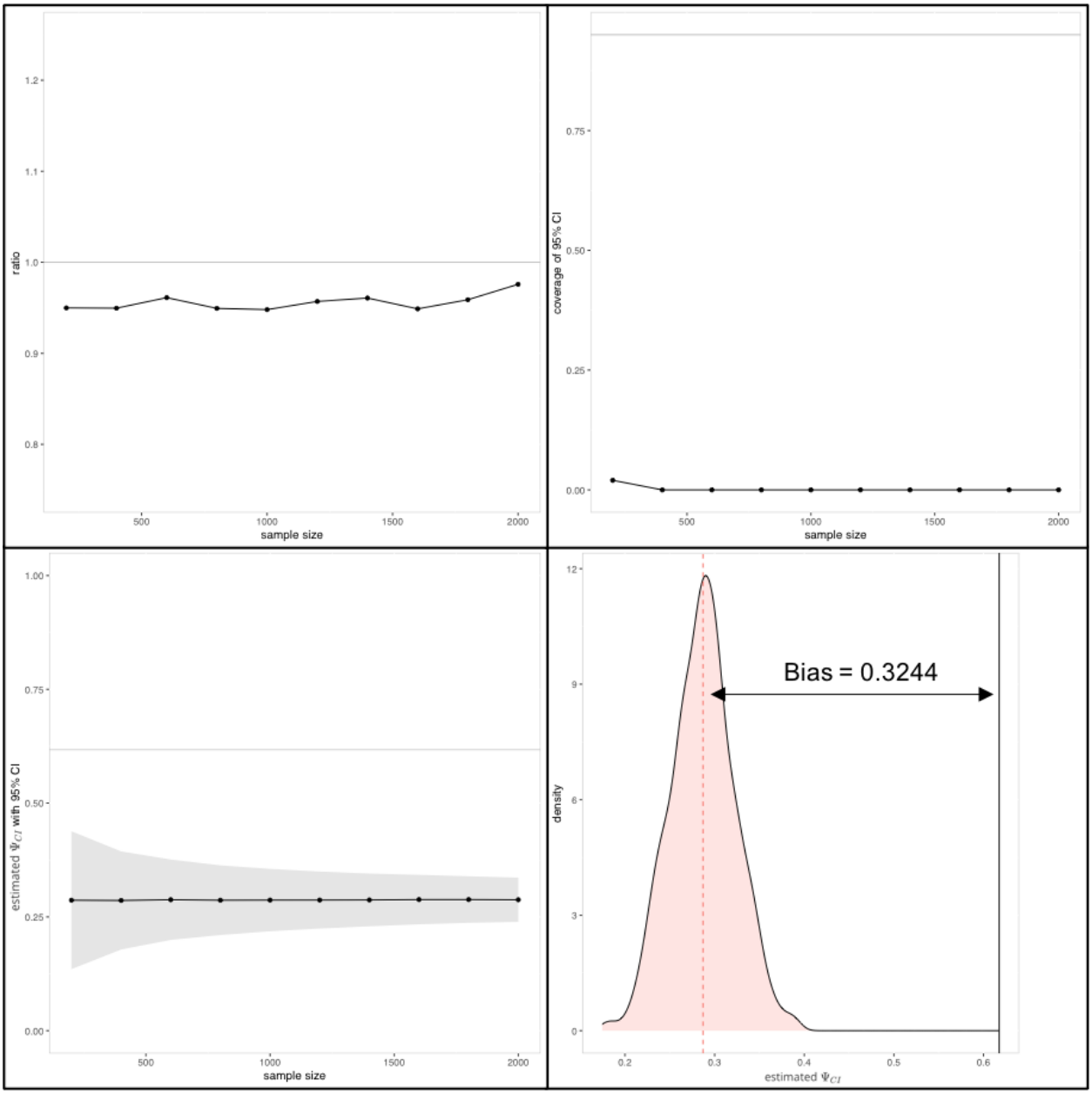


Figure 2.9: Simulation results when the conditional independence identification assumption is violated. **Upper left**: ratio of estimated variance over true variance vs sample size. The line is the ratio at each sample size, and the grey horizontal line is the true value; **Upper right**: coverage of 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve vs sample size. The line is the average coverage; **Lower left**: mean value of estimated $\psi$ and its 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve vs sample size. The line is the mean, and shaded area is the confidence interval and the grey horizontal line is the true value; **Lower right**: Distribution of 1000 estimates of $\psi$ for sample size of 1000, the vertical line is the true value, and the dashed line is the mean value.

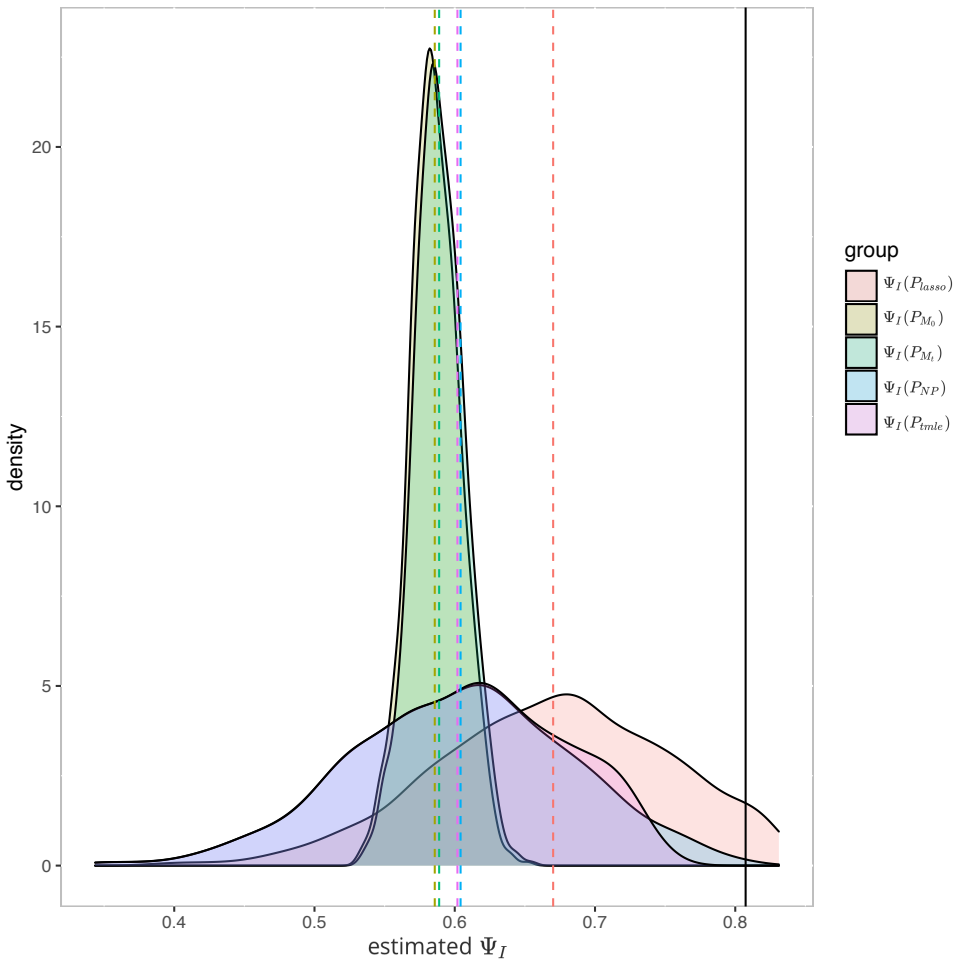


Figure 2.10: Distribution of 1000 estimates of $\Psi_I$ for sample size of 1000, the black vertical line is the true value, and the dashed vertical lines represent mean values for each estimator.

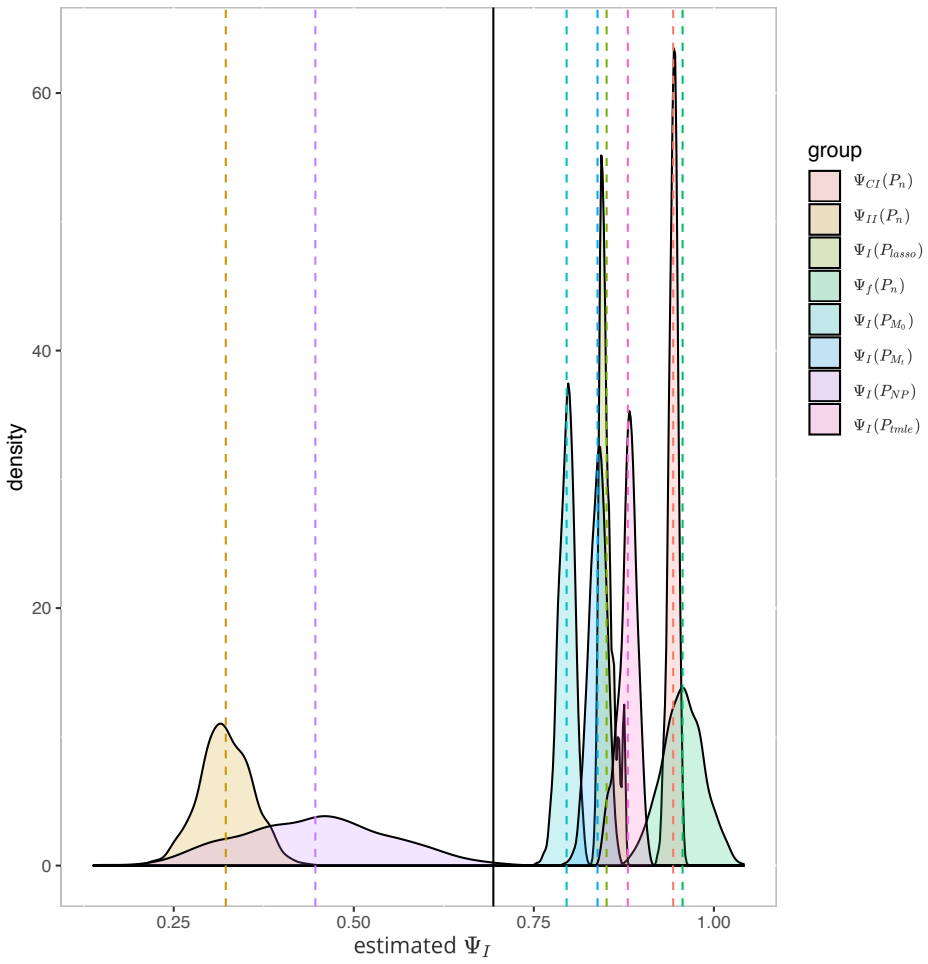


Figure 2.11: Distribution of 1000 estimates of $\psi$ for sample size of 1000, the black vertical line is the true value, and the dashed vertical lines represent mean values for each estimator.

| Assumption | Estimator | Average $\psi$ | Lower 95% CI | Upper 95% CI | Coverage(%) |
|---|---|---|---|---|---|
| Conditional independence | $\Psi_{CI}(P_n)$ | 0.9435 | 0.9313 | 0.9557 | 31.2 |
| Independence | $\Psi_{II}(P_n)$ | 0.3223 | 0.2472 | 0.3974 | 0 |
| Highest-way interaction equals zero (linear) | $\Psi_f(P_n)$ | 0.9565 | 0.8999 | 1.0132 | 0 |
| Highest-way interaction equals zero (log-linear) | $\Psi_I(P_{NP})$ | 0.4466 | 0.2519 | 0.6413 | 0 |
| Highest-way interaction equals zero (log-linear) | $\Psi_I(P_{lasso})$ | 0.8511 | 0.8005 | 0.9017 | 0 |
| Highest-way interaction equals zero (log-linear) | $\Psi_I(P_{tmle})$ | 0.8805 | 0.8385 | 0.9225 | 31.2 |
| Highest-way interaction equals zero (log-linear) | $\Psi_I(P_{M_0})$ | 0.7955 | 0.7621 | 0.8269 | 0 |
| Highest-way interaction equals zero (log-linear) | $\Psi_I(P_{M_t})$ | 0.8384 | 0.8074 | 0.8673 | 0 |

Table 2.9: Estimated $\hat{\psi}$ and 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve and its coverage for each estimator. True $\psi_0 = 0.6938$. Conditional independence assumption: sample $B_1$ and $B_2$ are independent conditional on $B_3$; Independence assumption: sample $B_1$ and $B_2$ are independent; K-way additive interaction equals zero assumption: 3-way interaction term in linear model equals zero; K-way multiplicative interaction equals zero assumption: 3-way interaction term in log-linear model equals zero.

| Estimator | Df | AIC | BIC |
|---|---|---|---|
| $\Psi_I(P_{M_0})$ | 5 | 743.631 | 753.447 |
| $\Psi_I(P_{M_t})$ | 3 | 391.757 | 411.388 |

Table 2.10: model fit statistics for$\Psi_I(P_{M_0})$ and $\Psi_I(P_{M_t})$ estimators.

## 2.7 Data Analysis

Schistosomiasis is an acute and chronic parasitic disease endemic to 78 countries worldwide. Estimates show that at least 229 million people required preventive treatment in 2018 [30]. In China, schistosomiasis is an ongoing public health challenge and the country has set ambitious goals of achieving nationwide transmission interruption [31]. To monitor the transmission of schistosomiasis, China operates three surveillance systems which can be considered as repeated samples of the underlying infected population. The first is a national surveillance system designed to monitor nationwide schistosomiasis prevalence ($S_1$), which covers 1% of communities in endemic provinces, conducting surveys every 6-9 years. The second is a sentinel system designed to provide longitudinal measures of disease prevalence and intensity in select communities ($S_2$), and conducts yearly surveys. The third system ($S_3$) comprises routine surveillance of all communities in endemic counties, with surveys conducted on a roughly three-year cycle [31]. In this application, we use case records for schistosomiasis from these three systems for a region in southwestern China [23]. The community, having a population of about 6000, reported a total of 302 cases in 2004: 112 in $S_1$; 294 in $S_2$; and 167 in $S_3$; with 202 cases appearing on more than one system register (figure 2.12). The three surveillance systems are not independent, and we apply estimators under the four constraints (highest-way interaction equals zero in log-linear model, highest-way interaction equals zero in linear model, independence, and conditional independence)to the data.

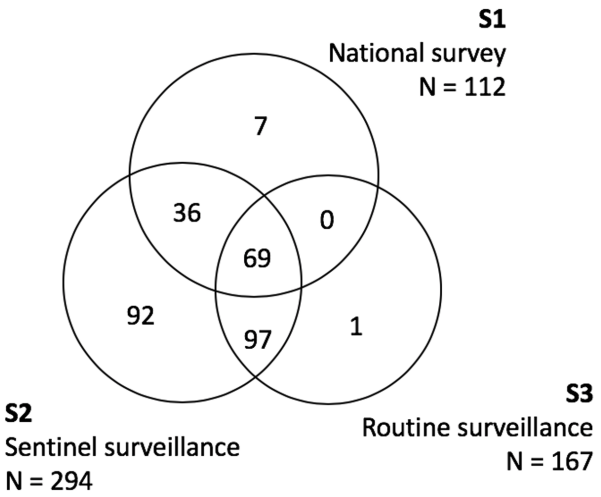


Figure 2.12: Schistosomiasis case frequencies among $S_1, S_2, S_3$ surveillance systems in a community (population $\sim 6000$) in southwestern China in 2004 [23]

The estimated $\hat{\psi}$ with its 95% asymptotic confidence intervals are shown in table 2.11. The highest estimate of $\psi$ is 0.9969 (with 95% CI: (0.8041, 0.9829)) under conditional in-

dependence assumption, and the lowest is 0.8935 (with 95% CI: (0.9965, 0.9972)) under the linear model with no highest-way interaction assumption (table 2.11). This analysis illustrates the heavy dependence of estimates of $\Psi$ on the arbitrary identification assumptions using real data.

One way to solve this problem is to hold certain identification assumptions true by design. In this case, we could make an effort to design the sentinel system $S_2$ and routine system $S_3$ to be independent of each other conditional on national system $S_1$. For example, let the sampling surveys in $S_2$ and $S_3$ be done independently, while each of them could borrow information from $S_1$. In so doing we can assume the conditional independence between the two samples and conduct the analysis correspondingly.

| Assumption | Estimator | Estimated $\psi$ | Lower 95% CI | Upper 95% CI |
|---|---|---|---|---|
| Highest-way interaction equals zero (log-linear) | $\Psi_I(P_{lasso})$ | 0.9212 | 0.8755 | 0.9669 |
| Highest-way interaction equals zero (log-linear) | $\Psi_I(P_{tmle})$ | 0.9428 | 0.9089 | 0.9768 |
| Highest-way interaction equals zero (log-linear) | $\Psi_I(P_{M_0})$ | 0.9321 | 0.8959 | 0.9627 |
| Highest-way interaction equals zero (log-linear) | $\Psi_I(P_{M_t})$ | 0.9934 | 0.9748 | 1 |
| Highest-way interaction equals zero (linear) | $\Psi_f(P_n)$ | 0.8935 | 0.8041 | 0.9829 |
| Independence | $\Psi_{II}(P_n)$ | 0.963 | 0.9326 | 0.9935 |
| Conditional independence | $\Psi_{CI}(P_n)$ | 0.9969 | 0.9965 | 0.9972 |

Table 2.11: Estimated $\hat{\psi}$ and 95% asymptotic confidence interval based on normal approximation with $\hat{\sigma}^2$ estimated from efficient influence curve and its coverage for each estimator. Conditional independence assumption: survey $S_1$ and $S_2$ are independent conditional on $S_3$; Independence assumption: survey $S_1$ and $S_2$ are independent; K-way additive interaction equals zero assumption: 3-way interaction term in linear model equals zero; K-way multiplicative interaction equals zero assumption: 3-way interaction term in log-linear model equals zero. The $\Psi_I(P_{NP})$ estimator under K-way multiplicative interaction term equals zero assumption is not defined due to existing empty cells in observed data.

| Estimator | Df | AIC | BIC |
|---|---|---|---|
| $\Psi_I(P_{M_0})$ | 5 | 329.383 | 336.804 |
| $\Psi_I(P_{M_t})$ | 3 | 60.139 | 74.980 |

Table 2.12: model fit statistics for$\Psi_I(P_{M_0})$ and $\Psi_I(P_{M_t})$ estimators.

## 2.8 Discussion

### Summary table

In this section we summarise all the estimators we proposed under linear and non-linear constraints as below.

| Identification assumption | Highest-way interaction term in linear model equals zero |
|---|---|
| Constraint | $E_{P_{B^*}} f(B^*) = \sum_b (-1)^{K+\sum_{k=1}^K b_k} P_{B^*,0}(b) = 0$ |
| Method | Plug-in |
| Target parameter | $\psi_f = \frac{f(0)}{f(0)-\sum_{b\neq 0} f(b)P(b)}$ |
| Efficient influence curve | $D_f^* = \frac{f(0)}{(f(0)-\sum_{b\neq 0} f(b)P(b))^2}[f(B) - \sum_{b\neq 0} f(b)P(b)]$ |

Table 2.13: Summary table for identification assumption: highest-way interaction term in linear model equals zero, where $f(b) = (-1)^{K+\sum_{k=1}^K b_k}$.

| Identification assumption | Independence between two samples |
|---|---|
| Constraint | $\Phi_{II}(P^*) \equiv \sum_b I(b(1:2)=(0,0))P^*(b)$<br>$-\sum_{b_1,b_2} I(b_1(1)=b_2(2)=0)P^*(b_1)P^*(b_2)$ |
| Method | Plug-in |
| Target parameter | $\Psi_{II}(P) = \frac{1-P(B(1)=0)-P(B(2)=0)+P(B(1:2)=(0,0))}{1-P(B(1)=0)-P(B(2)=0)+P(B(1)=0)P(B(2)=0)}$ |
| Efficient influence curve | $D^*_{\Phi_{II}}(P) = C_2(P)\{I(B(1)=0) - P(B(1)=0)\}$<br>$+ C_3(P)\{I(B(2)=0) - P(B(2)=0)$<br>$+ C_4(P)\{I(B(1)=B(2)=0) - P(B(1:2)=0\}$,<br>where $C_2(P) = \frac{1-P(B(1)=0)-P(B(2)=0)-P(B(1)=B(2)=0)}{(1-P(B(1)=0)-P(B(2)=0)-P(B(1)=0)P(B(2)=0))^2}P(B(2)=0)$<br>$C_3(P) = \frac{1-P(B(1)=0)-P(B(2)=0)-P(B(1)=B(2)=0)}{(1-P(B(1)=0)-P(B(2)=0)-P(B(1)=0)P(B(2)=0))^2}P(B(1)=0)$<br>$C_4(P) = -\frac{1}{1-P(B(1)=0)-P(B(2)=0)-P(B(1)=0)P(B(2)=0)}$ |

Table 2.14: Summary table for identification assumption: independence between two samples.

| Identification assumption | Conditional independence between two samples |
|---|---|
| Constraint | $\Phi_{CI,(j,m)} = P^*(B_j = 1 \mid B_1 = 0, \cdots, B_m = 0, \cdots, B_K = 0)$ - $P^*(B_j = 1 \mid B_1 = 0, \cdots, B_m = 1, \cdots, B_K = 0)$ |
| Method | Plug-in |
| Target parameter | $\Psi_{CI} = \frac{P(B_m=1,B_j=1,0,...,0)}{P(B_m=1,B_j=1,0,...,0)+P(B_m=0,B_j=1,0,...,0)P(B_m=1,B_j=0,0,...,0)}$ |
| Efficient influence curve | $D^*_{\Phi_{CI}}(P) = \frac{1}{C_5(P)}(C_6(P) - C_7(P) - C_8(P))$, where $C_5(P) = -\sum_{b_1\neq 0, b_2\neq 0}[\mathbb{I}(b_1(1) = b_2(2) = 0)P(b_1)P(b_2)$ $C_6(P) = P(B_m = 0, B_j = 1, 0, ..., 0)P(B_m = 0, B_j = 0, 0, ..., 0)$ $[\mathbb{I}(B_m = 1, B_j = 1, 0, ..., 0) - P(B_m = 1, B_j = 1, 0, ..., 0)]$ $C_7(P) = P(B_m = 1, B_j = 0, 0, ..., 0)P(B_m = 1, B_j = 1, 0, ..., 0)$ $[\mathbb{I}(B_m = 0, B_j = 1, 0, ..., 0) - P(B_m = 0, B_j = 1, 0, ..., 0)]$ $C_8(P) = P(B_m = 0, B_j = 1, 0, ..., 0)P(B_m = 1, B_j = 1, 0, ..., 0)$ $[\mathbb{I}(B_m = 1, B_j = 0, 0, ..., 0) - P(B_m = 1, B_j = 0, 0, ..., 0)]$ |

Table 2.15: Summary table for identification assumption: conditional independence between two samples.

| Identification assumption | Highest-way interaction term in log-linear model equals zero |
|---|---|
| Constraint | $\Phi_I(P^*) \equiv \sum_b (-1)^{1+\sum_{k=1}^K b_k} \log P_{B^*}(b) = 0$ |
| Method | Plug-in (NPMLE), undersmoothed lasso, TMLE based on lasso |
| Target parameter | $\Psi_I(P) = \frac{1}{1+\exp((-1)^{K+1}\sum_{b\neq 0} f(b)\log P(b))}$ |
| Efficient influence curve | $D^*_{\Phi_I}(P) = (-1)^K \Psi_I(P)(1 - \Psi_I(P))\left\{\frac{f(B)}{P(B)} + f(0)\right\}$ |

Table 2.16: Summary table for identification assumption: highest-way interaction term in log-linear model equals zero, where $f(b) = (-1)^{K+\sum_{k=1}^K b_k}$.

We developed a modern method to estimate population size based on capture-recapture designs with a minimal number of constraints or parametric assumptions. We provide the solutions, theoretical support, simulation study and sensitivity analysis for four identification assumptions: independence between two samples, conditional independence between two samples, no highest-way interaction in linear models, and no highest-way interaction in log-linear models. We also developed machine learning algorithms to solve the curse of dimensionality for high dimensional problems under the assumption of no highest-way interaction in log-linear model. Through our analysis, we found that whether the identification assumption holds true plays a vital role in the performance of estimation. When the assumption is violated, all estimators will be biased. This conclusion applies to models of all forms, parametric or non-parametric, simple plug-in estimators or complex machine-learning based estimators. Thus one should always ensure that the chosen identification assumption is known to be true by survey design, otherwise all the estimators will be unreliable. Under the circumstances where the identification assumptions hold true, the performance of our targeted maximum likelihood estimator, $\Psi_I(P_{tmle})$, is superior to $\Psi_I(P_{M_0})$(identical capture-probabilities, no highest-way interaction in log-linear model), $\Psi_I(P_{M_t})$(no interaction terms in log-linear model) and $\Psi_I(P_{NP})$(plugged-in, no highest-way interaction in log-linear model) estimators in several aspects: first, by making the least number of assumptions required for identifiability, the estimator $\Psi_I(P_{tmle})$ is more robust in empirical data analysis, as there will be no bias due to violations of parametric model assumptions. Second, the estimator $\Psi_I(P_{tmle})$ is based on a consistent undersmoothed lasso estimator. This property ensures the asymptotic efficiency of TMLE. Third, when there are empty cells, the estimator $\Psi_I(P_{tmle})$ solves the curse of dimensionality by correcting the bias introduced by the undersmoothed lasso estimator, and gives a more honest asymptotic confidence interval, wider that those from parametric models, and hence a higher coverage.

## 2.9 Appendix

In the appendix, we formally state the lemmas used in the context and provide the proofs. In section 2.9, we derive the target parameter $\Psi_{II}(P)$ under identification assumption that the first two samples $B_1, B_2$ are independent, given there are three samples in total. In section 2.9 we derive the efficient influence curve $D^*_{\Phi_{II}}(P)$ for $\Psi_{II}(P)$. In section 2.9, we state the lemma on how to derive the efficient influence curve under multidimensional constraint $\Phi_{III}$. In section 2.9, we formally state the target parameter $\Psi_{CI}(P)$ under identification assumption that the first two samples $B_1, B_2$ are independent conditional on the third samples. In section 2.9 we derive the efficient influence curve $D^*_{\Phi_{CI}}(P)$ for $\Psi_{CI}(P)$. In section 2.9 we prove the asymptotic efficiency of the TMLE.

### Target parameter under independence assumption

**Lemma 1.** *For constraint $\Phi_{II}$(equation 2.4), we have the target parameter $\Psi_{II}$ as:*

$$\Psi_{II}(P) = \frac{1 - P(B(1) = 0) - P(B(2) = 0) + P(B(1:2) = (0,0))}{1 - P(B(1) = 0) - P(B(2) = 0) + P(B(1) = 0)P(B(2) = 0)}$$

*Proof.* We provide a brief proof of lemma 1 when there are three samples. In this case, $\Phi_{II}(P^*) = 0$ is equivalent to

$$P^*(B^*(1:2) = (0,0)) = P^*(B(1) = 0)P^*(B^*(2) = 0). \tag{2.17}$$

When there are three samples, equation 2.17 can be expanded as:

$$\begin{aligned} P^*(0,0,1) + P^*(0,0,0) \quad = \quad & P^*(0,0,0)^2 + [P^*(0,1,0) + P^*(0,1,1) + P^*(0,0,1) \\ & + P^*(1,0,0) + P^*(1,0,1) + P^*(0,0,1)] \times P^*(0,0,0) \\ & + [P^*(0,1,0) + P^*(0,1,1) + P^*(0,0,1)] \\ & \times [P^*(1,0,0) + P^*(1,0,1) + P^*(0,0,1)] \end{aligned} \tag{2.18}$$

Denote $a^* = P^*(0,1,0) + P^*(0,1,1) + P^*(0,0,1)$, $b^* = P^*(1,0,0) + P^*(1,0,1) + P^*(0,0,1)$. Equation 2.18 can be written as:

$$0 = P^*(0,0,0)^2 + (a^* + b^* - 1) \times P^*(0,0,0) - P^*(0,0,1).$$

Plug in $\psi = 1 - P^*(0,0,0)$, we have

$$P(0,1,0) = \frac{P^*(0,1,0)}{\psi}, ..., P(0,0,1) = \frac{P^*(0,0,1)}{\psi}, a = \frac{a^*}{\psi}, b = \frac{b^*}{\psi}.$$

Thus equation 2.18 can be expressed as:

$$0 = (1 + ab - a - b)\psi^2 + (a + b - 1 - P(0,0,1))\psi. \tag{2.19}$$

Here, let $a_{II} = 1 + ab - a - b$, $b_{II} = a + b - 1 - P(0,0,1)$, from equation 2.19 we know $a_{II}\psi + b_{II} = 0$, thus we have

$$\begin{aligned}\psi &= -\frac{b_{II}}{a_{II}} \\ &= \frac{1 - P(B(1)=0) - P(B(2)=0) + P(B(1:2)=(0,0))}{1 + P(B(1)=0)P(B(2)=0) - P(B(1)=0) - P(B(2)=0)}\end{aligned}$$

where $P(B(1)=0) = P(0,1,0)+P(0,1,1)+P(0,0,1), P(B(2)=0) = P(1,0,0)+P(1,0,1)+P(0,0,1)$, and $P(B(1:2)=(0,0)) = P(0,0,1)$ □

## Efficient influence curve under independence assumption

**Lemma 2.** *For constraint $\Phi_{II} = 0$(equation 2.4), we have the efficient influence curve $D^*_{\Phi_{II}}(P)$ as:*

$$\begin{aligned}D^*_{\Phi_{II}}(P) &= \frac{\partial\psi}{\partial P(B(1)=0)} \times D^*_{\Phi_{II}}(P(B(1)=0)) \\ &\quad + \frac{\partial\psi}{\partial P(B(2)=0)} \times D^*_{\Phi_{II}}(P(B(2)=0)) \\ &\quad + \frac{\partial\psi}{\partial P(B(1:2)=(0,0))} \times D^*_{\Phi_{II}}(P(B(1:2)=(0,0))).\end{aligned} \tag{2.20}$$

*Proof.* By the delta method [32], the efficient influence curve of $\psi$ can be written as a function of each components' influence curve. The efficient influence curves of the three components are presented as follows:

$$\begin{aligned}D^*_{\Phi_{II}}(P(B(1)=0)) &= \mathbb{I}(B(1)=0) - P(B(1)=0). \\ D^*_{\Phi_{II}}(P(B(2)=0)) &= \mathbb{I}(B(2)=0) - P(B(2)=0). \\ D^*_{\Phi_{II}}(P(B(1:2)=(0,0))) &= \mathbb{I}(B(1:2)=(0,0)) - P(B(1:2)=(0,0)).\end{aligned}$$

Therefore, we only need to calculate the derivatives to get $D^*_{\Phi_{II}}(P)$, and the three parts of derivatives are given by

$$\frac{\partial\psi}{\partial P(B(1)=0)} = \frac{(1-P(B(2)=0))(P(B(1:2)=(0,0)) - P(B(1)=0))}{(1-P(B(1)=0) - P(B(2)=0) + P(B(1)=0)P(B(2)=0))^2}.$$

$$\frac{\partial\psi}{\partial P(B(2)=0)} = \frac{(1-P(B(1)=0))(P(B(1:2)=(0,0)) - P(B(1)=0))}{(1-P(B(1)=0) - P(B(2)=0) + P(B(1)=0)P(B(2)=0))^2}.$$

$$\frac{\partial\psi}{\partial P(B(1:2)=(0,0))} = \frac{1}{1-P(B(1)=0) - P(B(2)=0) + P(B(1)=0)P(B(2)=0)}.$$

Plug each part into equation 2.20, we have

$$\begin{aligned} D^*_{\Phi_{II}}(P) &= \frac{(1-P(B(2)=0))(P(B(1:2)=(0,0))-P(B(2)=0))}{(1-P(B(1)=0)-P(B(2)=0)+P(B(1)=0)P(B(2)=0))^2} \\ &\quad [\mathbb{I}(B(1)=0)-P(B(1)=0)] \\ &\quad + \frac{(1-P(B(1)=0))(P(B(1:2)=(0,0))-P(B(1)=0))}{(1-P(B(1)=0)-P(B(2)=0)+P(B(1)=0)P(B(2)=0))^2} \\ &\quad [\mathbb{I}(B(2)=0)-P(B(2)=0)] \\ &\quad + \frac{1}{1-P(B(1)=0)-P(B(2)=0)+P(B(1)=0)P(B(2)=0)} \\ &\quad [\mathbb{I}(B(1:2)=(0,0))-P(B(1:2)=(0,0))]. \end{aligned}$$

□

## Efficient influence curve under multidimensional constraint

**Lemma 3.** *Consider a model $M \equiv \{P \in M_1 : \Phi(P) = 0\}$ defined by an initial larger model $M_1$ and multivariate constraint function $\Phi : M_1 \to I\!R^K$. Suppose that $\Phi : M_1 \to I\!R^K$ is path-wise differentiable at $P$ with efficient influence curve $D^*_\Phi(P)$ for all $P \in M_1$. Let $T_1(P)$ be the tangent space at $P$ for model $M_1$, and let $\Pi_{T_1} : L^2_0(P) \to T_1(P)$ be the projection operator onto $T_1(P)$. The tangent space at $P$ for model $M$ is given by:*

$$T(P) = \{S \in T_1(P) : S \perp D^*_\Phi(P)\}.$$

*The projection onto $T(P)$ is given by:*

$$\Pi_T(S) = \Pi_T(S) - \Pi_{D^*_\Phi}(\Pi_T(S)),$$

*where $\Pi_{D^*_\Phi}$ is the projection operator on the $K$-dimensional subspace of $T_1(P)$ spanned by the components of $D^*_\Phi(P)$. The latter projection is given by the formula:*

$$\Pi_{D^*_\Phi}(S) = E(SD^*_\Phi(P)^\top)E(D^*_\Phi(P)D^*_\Phi(P)^\top)^{-1}D^*_\Phi(P).$$

## Target parameter under conditional independence assumption

**Lemma 4.** *For constraint $\Phi_{CI} = 0$ (equation 2.5), we have the target parameter $\Psi_{CI}$ as*

$$\Psi_{CI} = \frac{P(B_m = 1, B_j = 1, 0, ..., 0)}{C_0(P)}.$$

*where*

$$\begin{aligned} C_0(P) &= P(B_m = 1, B_j = 1, 0, ..., 0) \\ &\quad + P(B_m = 0, B_j = 1, 0, ..., 0)P(B_m = 1, B_j = 0, 0, ..., 0). \end{aligned}$$

*Proof.* The constrain $\Phi_{CI} = 0$(equation 2.5) can be written as

$$\begin{aligned} &P^*(B_j = 1|B_1 = 0, \cdots, B_m = 0, \cdots, B_K = 0) \quad = \\ &P^*(B_j = 1|B_1 = 0, \cdots, B_m = 1, \cdots, B_K = 0). \end{aligned}$$

This equation is equivalent to

$$\begin{aligned} P^*(0, \ldots, 0) \quad = \quad & \frac{P^*(B_m = 0, B_j = 1, 0, \ldots, 0)}{P^*(B_m = 1, B_j = 1, 0, \ldots, 0)} \\ & \times P^*(B_m = 1, B_j = 0, 0, \ldots, 0). \end{aligned} \tag{2.21}$$

As $\Psi_{CI} \equiv 1 - P^*(0, \ldots, 0)$, we have

$$P^*(B_m = 1, B_j = 0, 0, \ldots, 0) = P(B_m = 1, B_j = 0, 0, \ldots, 0)\Psi_{CI}.$$

$$P^*(B_m = 0, B_j = 1, 0, \ldots, 0) = P(B_m = 0, B_j = 1, 0, \ldots, 0)\Psi_{CI}.$$

$$P^*(B_m = 1, B_j = 1, 0, \ldots, 0) = P(B_m = 1, B_j = 1, 0, \ldots, 0)\Psi_{CI}.$$

Therefore, equation 2.21 is equivalent to

$$\Psi_{CI} = \frac{P(B_m = 1, B_j = 1, 0, \ldots, 0)}{P(B_m = 1, B_j = 1, 0, \ldots, 0) + C_9}.$$

where

$$C_9 = P(B_m = 0, B_j = 1, 0, \ldots, 0)P(B_m = 1, B_j = 0, 0, \ldots, 0).$$

□

## Efficient influence curve under conditional independence assumption

**Lemma 5.** *For constraint $\Phi_{CI} = 0$ (equation 2.5), we have the efficient influence curve $D^*_{\Phi_{CI}}(P)$ as*

$$D^*_{\Phi_{CI}}(P) \quad = \quad \frac{1}{C_5(P)}(C_6(P) - C_7(P) - C_8(P))$$

*where*

$$\begin{aligned} C_5(P) \quad &= \quad - \sum_{b_1 \neq 0, b_2 \neq 0} I(b_1(1) = b_2(2) = 0)P(b_1)P(b_2). \\ C_6(P) \quad &= \quad P(B_m = 0, B_j = 1, 0, ..., 0)P(B_m = 0, B_j = 0, 0, ..., 0) \\ &\quad [\mathbb{I}(B_m = 1, B_j = 1, 0, ..., 0) - P(B_m = 1, B_j = 1, 0, ..., 0)]. \\ C_7(P) \quad &= \quad P(B_m = 1, B_j = 0, 0, ..., 0)P(B_m = 1, B_j = 1, 0, ..., 0) \\ &\quad [\mathbb{I}(B_m = 0, B_j = 1, 0, ..., 0) - P(B_m = 0, B_j = 1, 0, ..., 0)]. \\ C_8(P) \quad &= \quad P(B_m = 0, B_j = 1, 0, ..., 0)P(B_m = 1, B_j = 1, 0, ..., 0) \\ &\quad [\mathbb{I}(B_m = 1, B_j = 0, 0, ..., 0) - P(B_m = 1, B_j = 0, 0, ..., 0)]. \end{aligned}$$

*Proof.* By the delta method [32], the efficient influence curve of $\Psi_{CI}$ can be written as a function of each components' influence curve. The efficient influence curves of the three components are presented as follows

$$\begin{aligned} D^*_{\Phi_{CI}}(P(B_m=1,B_j=1,0,\ldots,0)) &= \mathbb{I}(B_m=1,B_j=1,0,\ldots,0) \\ &\quad -P(B_m=1,B_j=1,0,\ldots,0). \qquad (2.22) \\ D^*_{\Phi_{CI}}(P(B_m=0,B_j=1,0,\ldots,0)) &= \mathbb{I}(B_m=0,B_j=1,0,\ldots,0) \\ &\quad -P(B_m=0,B_j=1,0,\ldots,0). \qquad (2.23) \\ D^*_{\Phi_{CI}}(P(B_m=1,B_j=0,0,\ldots,0)) &= \mathbb{I}(B_m=1,B_j=0,0,\ldots,0) \\ &\quad -P(B_m=1,B_j=0,0,\ldots,0)). \qquad (2.24) \end{aligned}$$

Therefore, we only need to calculate the derivatives to get $D^*_{\Phi_{CI}}(P)$, and the three parts of derivatives are given by

$$\begin{aligned} \frac{\partial\psi}{\partial P(B_m=1,B_j=1,0,\ldots,0)} &= \frac{P(B_m=0,B_j=1,0,\ldots,0)}{C_4} \\ &\quad \times P(B_m=1,B_j=0,0,\ldots,0). \\ \frac{\partial\psi}{\partial P(B_m=0,B_j=1,0,\ldots,0)} &= -\frac{P(B_m=1,B_j=1,0,\ldots,0)}{C_4} \\ &\quad \times P(B_m=1,B_j=0,0,\ldots,0). \\ \frac{\partial\psi}{\partial P(B_m=1,B_j=0,0,\ldots,0)} &= -\frac{P(B_m=1,B_j=1,0,\ldots,0)}{C_4} \\ &\quad \times P(B_m=1,B_j=1,0,\ldots,0). \end{aligned}$$

where

$$\begin{aligned} C_4 &= [P(B_m=1,B_j=1,0,\ldots,0)+P(B_m=0,B_j=1,0,\ldots,0) \\ &\quad \times P(B_m=1,B_j=0,0,\ldots,0)]^2. \end{aligned}$$

Plug each part into equation 2.24, we have that lemma 5 is true. □

## Asymptotic efficiency of TMLE

In this section we prove the following theorem 1 establishing asymptotic efficiency of the TMLE. The empirical NPMLE is also efficient since asymptotically all cells are filled up. And TMLE is just a finite sample improvement and asymptotically TMLE acts as the empirical NPMLE. The estimators $P_n*$ will be like parametric model MLE and thus converge at rate $o_P(1/\sqrt{n})$.

**Theorem 1.** *Consider the TMLE $\Psi_I(P^*_n)$ of $\Psi_I(P_0)$ defined above, satisfying $P_nD^*_{\Phi_I}(P^*_n) = o_P(1/\sqrt{n})$. We assume $P_0(b) > 0$ for all $b \neq 0$. If $P^*_n - P_0 = o_P(n^{-1/4})$, then $\Psi_I(P^*_n)$ is an*

*asymptotically efficient estimator of* $\Psi_I(P_0)$*:*

$$\Psi_I(P_n^*) - \Psi_I(P_0) = (P_n - P_0)D_{\Phi_I}^*(P_0) + o_P(1/\sqrt{n}).$$

*Proof.* Define $f_{I,0}(b) = f_I(b) + f_I(0)$.Note that

$$P_0 D_{\Phi_I}^*(P) = (-1)^K \psi_I(1-\psi_I) \sum_{b\neq 0} f_{I,0}(b) \frac{(P_0 - P)(b)}{P(b)}.$$

Consider the second order Taylor expansion of $\Psi_I(P_0)$ at $P$:

$$\Psi_I(P_0) - \Psi_I(P) = d\Psi_I(P)(P_0 - P) + R_2(P, P_0),$$

where

$$\begin{aligned} d\Psi_I(P)(h) &= \frac{d}{d\epsilon}\Psi_I(P+\epsilon h)\Big|_{\epsilon=0} \\ &= (-1)^K \Psi_I(P)(1-\Psi_I(P)) \sum_{b\neq 0} \frac{f_I(b)h(b)}{P(b)}, \end{aligned}$$

$$\begin{aligned} R_2(P, P_0) &= \frac{1}{2}\frac{d^2\Psi_I(P+\epsilon h)}{d\epsilon^2}(P_0 - P)^2\Big|_{\epsilon=0} + o_P((P_0 - P)^2) \\ &= \frac{1}{2}\Psi_I(P)(1-\Psi_I(P))(P_0 - P)^2(b) \times \\ &\quad [(1-2\Psi_I(P))(\sum_{b\neq 0} \frac{f_I(b)h(b)}{P(b)})^2 + (-1)^{K+1} \sum_{b\neq 0} \frac{f_I(b)h(b)^2}{P(b)^2}] \\ &\quad + o_P((P_0 - P)^2(b)) \end{aligned} \tag{2.25}$$

From equation 2.25 we know that $R_2(P, P_0)$ is a second order term involving square differences $(P_0 - P)^2(b)$ for $b \neq 0$. Thus, we observe that

$$P_0 D_{\Phi_I}^*(P) = d\Psi_I(P)(P_0 - P).$$

This proves that

$$P_0 D_{\Phi_I}^*(P) = \Psi_I(P_0) - \Psi_I(P) - R_2(P, P_0).$$

We can apply this identity to $P_n^*$ so that we obtain $P_0 D_{\Phi_I}^*(P_n^*) = \Psi_I(P_0) - \Psi_I(P_n^*) - R_2(P_n^*, P_0)$. Combining this identity with $P_n^* D_{\Phi_I}^*(P_n^*) = 0$ yields:

$$\Psi_I(P_n^*) - \Psi_I(P_0) = (P_n - P_0)D_{\Phi_I}^*(P_n^*) + R_2(P_n^*, P_0).$$

We will assume that $P_n^*$ is consistent for $P_0$ and $P_0(b) > 0$ for all $b \neq 0$, so that it follows that $D^*_{\Phi_I}(P_n^*)$ is uniformly bounded by a $M < \infty$ with probability tending to 1, an that it falls in a $P_0$-Donsker class (dimension is finite $2^K - 2$), and $P_0\{D^*_{\Phi_I}(P_n^*) - D^*_{\Phi_I}(P_0)\}^2 \to 0$ in probability as $n \to \infty$. By empirical process theory, it now follows that

$$(P_n - P_0)D^*_{\Phi_I}(P_n^*) = (P_n - P_0)D^*_{\Phi_I}(P_0) + o_P(1/\sqrt{n}).$$

We also note that $R_2(P_n^*, P_0)$ has denominators that are bounded away from zero, so that $R_2(P_n^*, P_0) = o_P(1/\sqrt{n})$ if $P_n^* - P_0 = o_P(n^{-1/4})$ (e.g., Euclidean norm). Thus theorem 1 is proved. □

# Chapter 3

# Targeted learning in assessing the health care program performance

## 3.1 Introduction

### Background of the health care program

In Mexico, type 2 diabetes (T2D) is a major public health concern. The prevalence of this condition is above 9.4% in the adult population and increasing [33]. T2D is a chronic disease characterized by a progressive loss of $\beta$-cell insulin secretion and frequent insulin resistance [34]. In poorly controlled patients, the chronic hyperglycemia causes damage of multiple organ systems and development of micro- and macrovascular complications. The manifestations of microvascular complications are nephropathy, retinopathy and neuropathy. Macrovascular complications are coronary artery disease, peripheral arterial disease, and stroke. These complications are accountable for most of the morbidity, hospitalizations, and deaths that occur in patients with diabetes mellitus [35, 36]. A recent meta-analysis of 28 randomized trials that included 34,912 T2D patients found that targeting intensive glycemic control (HbA1C $< 7\%$) reduces the risk of microvascular complications, compared with conventional glycemic control; yet, it also increases the risk of hypoglycemia and did not show significant differences for all-cause and cardiovascular mortality [37].

The Mexican Institute of Social Security (Spanish acronym IMSS), is the most extensive health system in Mexico with nearly 65 million affiliates provides care to approximately 3.8 million T2D adult patients. The growing demand and healthcare requirements of T2D pose a heavy burden for family medicine clinics (FMCs), the frontline of IMSS healthcare. T2D patients are the second cause of consultation at FMCs, and those with acute and chronic complications, including comorbidities (i.e., hypertension) are among the top ambulatory and emergency consultations and hospital admissions [38]. Furthermore, T2D has substantial economic consequences, since in 2016, diabetes expenditures alone accounted for 2.5 billion US dollars [39].

T2D is a complex chronic condition that requires multidisciplinary healthcare and strict

patient's adherence to reduce the risk of acute and chronic complications. The primary goal of T2D treatment is to reach glucose control (glycated hemoglobin -HbA1C- below 7%). Conventionally, IMSS FMC consultations and follow-ups for T2D have been provided by a family doctor including physical examination, laboratory tests (i.e., blood glucose) prescription of treatment and self-care counseling. The family doctor refers patients to the dietitian, social worker, ophthalmologist or other specialists for a consultation, but the frequency of referrals and waiting time to receive multidisciplinary care might last several weeks or months due to the limited supply of these specialists and the increasing demand of patients with T2D. An analysis of the electronic health records of 25,130 T2D patients found that only 13% were referred to an ophthalmologist, 3.9% received nutritional counseling, and 23% had HbA1c $< 7\%$ (or plasma glucose $\leq 130$ mg/dl) [40]. Though there are specific clinical guidelines for T2D treatment, care is irregular and uncoordinated [41]. Evaluations of patient outcomes of FMCs at IMSS revealed less than 30% of T2D patients achieved HbA1c below 7% [42, 43, 44].

The need to improve health outcomes of T2D patients prompted IMSS to design and launch the DIABETIMSS program in 2008. DIABETIMSS is a comprehensive model of care that fulfills the Chronic Care Model attributes [45, 45]. The building block of DIABETIMSS is a multidisciplinary team (medical doctor, nurse, psychologist, dietitian, dentist, and social worker) that delivers coordinated and comprehensive healthcare. In addition to regular consultations with the team, T2D patients receive individual, family and group education on self-care and prevention of complications. Only T2D patients with less than 10 years after diagnosis and without severe chronic complications are eligible to enter DIABETIMSS. The primary goal of DIABETIMSS focuses on improving patient's self-care and achieving glycemic and metabolic control (reducing high blood pressure, cholesterol levels, and excess body fat, among others). Ultimately, DIABETIMSS care is expected to avert acute complications, reduce demand for emergency services and hospitalizations and delay the progression of organ damage.

The program has expanded gradually. Currently, $\sim$ 91,000 patients attend 136 DIABETIMSS program modules distributed throughout the country. DIABETIMSS introduced healthcare delivery changes for which an effectiveness evaluation is worthwhile. Previous evaluations of DIABETIMSS reported improvements in patient self-care and reductions in blood glucose levels. However, small samples and lack of a control group limit drawing robust conclusions [46, 47, 48]. In fact, in complex health systems, such as IMSS, it might not be possible to evaluate a new program by design, as it can be impractical to randomize the initiation of the program across different clinics for logistic and organizational reasons; therefore, to evaluate the impact of a program one often must rely on observational data.

## Background of the statistical methods

A new trend of statistical approaches including targeted Learning [2] and Super Learner [7] have been developed to use routine health data (e.g., electronic health records) to adjust for confounding and produce robust results that estimate parameters, such as the average

treatment effect (ATE). The motivation for these methods are the desire to have a combination of study design and analytical methods that can use observational data, but produce results akin to the robustness of a randomized control trial. Given that is often impractical to randomize the initiation of a program across different clinics, one often must rely on observational data to assess program impacts. Because controlling the experiment by design is no longer possible, one must use a combination of rich data that contains potentially high-dimensional patient/clinic variables and statistical methods to tease apart the impacts of programs from other potential competing causes of health outcomes.

Traditionally, standard parametric regression methods (e.g., logistic regression) were used to adjust for confounding factors in order to derive estimates of the associations. The main motivation for such methods was a combination of availability (available in standard statistical software) and interpretability (produce coefficients convenient interpretations of relative importance, such as odds ratios). However, because such models are inevitably biased (typically incorrect to assume the relationship of health outcome and predictor variables has simple form), the results could not be trusted the same as those from an "equivalent" randomized trial. However, the recently manifest combination of new modes of collecting data (e.g., electronic health record data, new forms of biomarkers of disease), accessible machine learning algorithms and the theory to develop combine the two (machine learning for estimation of causal impacts) allows the possibility of robustly evaluating health programs. This chapter demonstrates the practicality of leveraging these new forms of information and methods towards robust systems of program evaluation.

One of the main goals in this chapter was to set-up an efficient and robust data analytical stream that could estimate impacts of the program (or similar program evaluation analyses) using the most modern statistical methodology that relies on as few arbitrary assumptions as possible. Recent methodological developments have created the opportunity for analytic streams that take very little user-input, yet produce results that with trustworthy statistical inference (e.g., confidence intervals with the desired properties). This chapters uses the super learner [7] and causal inference [49] to define quantities of interest that are directly related to desired evaluation statistics, and targeted learning [2] to leverage these two towards estimation of these quantities. In addition, we have used methods that not only estimate the population impact of the program, but exploratory methods that can highlight among which clinics and patient sub-groups the program is working best. The latter are methods being developed within the context of precision medicine and precision public health. Ultimately, with sufficient data of high quality, one can even develop treatment rules to target patients that are most likely to benefit from the program in a context of limited resources. The ultimate goal is to create the basis for a dashboard analytical framework that can be applied across large medical systems for evaluating programs/treatments using sophisticated machine learning technology but with simple interfaces for non-technical users. Evaluation of the DIABETIMSS program provides a unique opportunity to both evaluate the success of the program in addressing the progression of diabetes, but also to see how the combination of detailed data collection and statistical algorithms move towards this goal, with the target of improving care.

## 3.2 Methodology

### Data

The data used were generated by IMSS. We analyzed the information from 11 family medicine clinics (FMC) located in the Mexico City and in the state of Mexico. We included in the study 6 clinics that implemented DIABETIMSS between 2008 and 2011 and 5 without DIABETIMSS program (those with conventional model of care). We selected by convenience three DIABETIMSS FMCs from the Mexico City and three from the State of Mexico, including in the study clinics with complete 2011-2016 laboratory databases (not all FMC clinics had laboratory databases for the above-mentioned period). The control clinics were randomly selected from the list of FMCs without DIABETIMSS program, choosing from those within the same geographic area and with the similar number of examining rooms (1014, 1524 and $\geq 25$). The sources of information were the IMSS electronic health records and clinical laboratory databases [39].

Besides being a high volume health care system, IMSS has an established Electronic Health Record system, which in turn is a component of the IMSS information system that includes administrative data (i.e., personnel information, inventories, characteristics of clinical settings), affiliation data (i.e.,socioeconomic information of affiliates and beneficiaries) and financial data among others.

### Statistical Methodology

We used both standard regression and targeted learning methods to evaluate the impact of the program. First, we performed simple bi-variate analyses, looking at the association of the indicator of glucose control (via the HbA1c indicator) and each of the predictor variables (the HbA1c indicator is based either directly on HbA1c levels or inferred from fasting glucose levels if data on HbA1c was lacking). For continuous variables, we use generalized versions of the t-test (that account for correlation of repeated outcome measures on a subject) or contingency table analysis for categorical predictors, via the generalized estimating equations approach [50]. We also used bar graphs to demonstrate the distribution of diabetes complications across those in and out of the DIABETIMSS program.

#### Parameters of Interest

First, we formally define the data structure. We assume the data are independent individuals, $i = 1, \cdots, m$, with repeated observations, $j = 1, ..., n_i$, that is we allow for some subjects to have fewer than the total possible number of times, 5. Then the data on each person can be represented by:

$$O_i \equiv (R_{ij}, T_{ij}, W_{ij}, A_{ij}, Y_{ij}, j = 1, \cdots, n_i), \tag{3.1}$$

where $R_{ij}$ is the specific clinic, $T_{ij}$ is the year of the program and measurement of outcome, $A_{ij}$ is the indicator of the DIABETIMSS program ($1 = yes$), $W_{ij}$ is the set of confounders that can include measurements made in past years, and $Y_{ij}$ is the indicator of glucose control.

In this analysis, we treat the data like a serial cross-sectional study, so define the observed data for an individual at time $T_{ij} = t$. In this case, $P_0(O(T_{ij}) = t)$ is the joint data-generating distribution of the data of observations made at time $t$. We estimate the association parameter based upon causal inference, separately by clinic, $R_{ij} = r$, but we average the impact over the years ($T_{ij} = t$) of the study. We define the yearly parameter of interest as:

$$\Psi(P)(r) = E\{E(Y_1 - Y_0 \mid T_{ij} = t, R_{ij} = r)\} \overset{assumptions}{=} \Psi(P_0)(r) \tag{3.2}$$

$$= E\{E\left[E(Y_{ij}|A_{ij} = 1, T_{ij} = t, W_{ij}, R_{ij} = r) - E(Y_{ij}|A_{ij} = 0, T_{ij} = t, W_{ij}, R_{ij} = r)\right]\}. \tag{3.3}$$

The left equation ($E(Y_1 - Y_0 \mid T_{ij} = t, R_{ij} = r)$ in (3.2) is a causal mean difference, where $Y_a$ is the so-called *counterfactual* outcome had the patient, possibly contrary to fact, had level $a$ of the intervention (in our case $A(t) = 1 \rightarrow$ patient is on the DIABETIMSS, 0 if not in program). Our parameter averages the annual association over the years of the program ($T = 2012, \ldots, 2016$). The statistical association parameter (3.3) can be thought of as the average of the stratified average differences in the proportion of subjects that achieve glucose control in the DIABETIMSS program versus those not in the program. Given that the outcome of interest in this case is binary (glucose control yes/no), the adjusted mean differences (3.2) can be thought of adjusted risk differences. If the assumptions are met, then the numbers returned by the procedure can be interpreted as the difference in the proportion of subjects with glucose control in the DIABETIMSS program versus those not enrolled.

In addition to the clinic specific estimates, we also reported estimates pooled across all the DIABETIMSS clinics:

$$\begin{aligned} \Psi_{pooled}(P_0) &= E_R\left(E\{E\left[E(Y_{ij}|A_{ij} = 1, T_{ij} = t, W_{ij}, R_{ij} = r)\right.\right. \\ &- E(Y_{ij}|A_{ij} = 0, T_{ij} = t, W_{ij}, R_{ij} = r)\}, \end{aligned}$$

(3.5)

where we simply take the mean now over the distribution of the population across clinics by adding the external expectation operator $E_R$ to (3.3) to get (3.4). Thus, the main parameters of interest can be interpreted as adjusted means, where the associations are adjusted for a set of potential confounders. We define the parameters of interest in a way that does not rely on a parametric regression model, e.g.,

$$E(Y_{ij}|A_{ij}, W_{ij}, T_{ij} = t, R_{ij} = r) = \beta_0^{r,t} + \beta_1^{r,t} A_{ij} + \beta_2^{r,t} W_{ij} \tag{3.6}$$

so that we can estimate meaningful potentially causal parameters without resorting to misspecified parametric models, such as (3.6). In essence, we try to get back to the statistical assumptions of a typical randomized clinical trial, where one can estimate similar parameters in a nonparametric model, that is, get the inferences as close to a randomized trial as possible.

**Comparisons to Traditional Statistical Techniques**

To note, if one takes (3.6) as the true model, then for a fixed $r, t$,

$$\begin{aligned}\beta_1^{r,t} &= E\left[E(Y_{ij}|A_{ij}=1, T_{ij}=t, W_{ij}, R_{ij}=r)\right. \\ &- \left. E(Y_{ij}|A_{ij}=0, T_{ij}=t, W_{ij}, R_{ij}=r)\right],\end{aligned}$$

so that the parameter we propose estimating is, in these very special cases, equivalent to a coefficient in multivariate regression model. However, in our case, we derive unbiased estimates whether or not the underlying true model is of a specific parametric form.

## Estimation

We estimate the above parameters of interest using three basic methods: 1) simple unadjusted mean differences, 2) adjusted via standard logistic regression and 3) adjusted via machine learning. For 3), we use an approach that first optimizes the prediction model (dropping the $ij$ for now), $Q(A, W, t, r) \equiv E(Y|A, W, T=t, R=r)$ and then targets that initial prediction fit towards estimation of the target parameter using Targeted Learning [2]. Given we do not know the true data-generating distribution, we can not use the data directly to assess the relative performance of the more advanced machine learning methods with simpler and more traditional approaches. Thus, we also use the data to flexibly estimate the relevant parts of the data-generating distribution. Once this is done, we forward simulate from these (where now the data-generating distribution of thus parameter of interest is known) and compare the performance of the more advanced estimators with traditional approaches (see Simulation section below).

### SuperLearning

For estimating the prediction function, $Q$, we use the ensemble machine learning algorithm called SuperLearner (SL; [7]). The SL algorithm uses cross-validation to avoid over-fitting (choosing a overly complex model). The SL algorithm works by taking a weighted average of the included prediction algorithms (learners), that optimizes the cross-validated fit of the resulting prediction function. To ensure optimal performance in a wide-variety of situations, SL should include both very simple learners (such as standard regression models) as well as more flexible, machine learning algorithms. To cover a wide landscape of potential prediction functions, we included the following library of learners in our SL fit:

1. mean: average of outcome given predictors (constant model),

2. earth: multivariate adaptive regression splines [51], which provides flexible fits via use of linear splines and adaptively chosen knots,

3. xgboost: extreme gradient boosting, which produces a fit that can be interpreted as a weighted average of regression tree fits [52]

4. standard logistic regression,

5. stepwise hierarchical logistic regression allowing for 2-way interactions,

6. ranger: computational fast implementation of so called random forests [6].

In our analyses that pool across clinics and require cross-validation (CV), we clustered form of CV that keeps all observations in the same clinic, $R$, within the same validation sample. This is done again to avoid over-fitting, which can occur when clustered data are separated in different validation samples. For all analyses, we used the statistical programming language R [53]. For the SL fits, we used the SuperLearner R package [54]. The package not only returns an estimate of the prediction function $Q$, but also provides estimates of the performance of the prediction model. In our case, we also report resulting cross-validated receiver-operator curves and associated estimated area-under-the-curves (AUC's) [55].

**Substitution Estimator**

We use substitution estimators to estimate the parameters discussed above (the adjusted risk differences) which empirically estimates the outer expectations in the parameters of interest by plugging an estimate, $\hat{Q}$, for the corresponding conditional mean in 3.3. Thus, the estimator has the form:

$$\Psi(P_n)(r) = \frac{1}{n_r} \sum_{t=2012}^{2016} \sum_{i=1}^{m} \sum_{j=1}^{n_i} I(R_{ij} = r) * I(T_{ij} = t) * (\hat{Q}(1, W_{ij}, t, r) - \hat{Q}(0, W_{ij}, t, r)), \quad (3.7)$$

where $n_r = \sum_{i=1}^{m} \sum_{j=1}^{n_i} I(R_{ij} = r)$. We fit three estimators of $\hat{Q}$ that result in three estimators of $\Psi$: unadjusted, maximum likelihood model (logistic regression), and targeted maximum likelihood (tmle and SL). For the unadjusted, we simply compare the proportions of glucose control among those in and out of DIABTIMSS program in a clinic in a particular year:

$$\Psi(P_n)(r) = \frac{1}{n_r} \sum_{t} \sum_{i} \sum_{j} I(R_{ij} = r) * I(T_{ij} = t) * (\hat{Q}^0(1, t, r) - \hat{Q}^0(0, t, r)), \quad (3.8)$$

where $\hat{Q}^0(1, t, r)$ is simply the proportion of observations with $Y_{ij} = 1$ among all observations in clinic $r$ ($R_{ij} = r$) within year $T_{ij} = t$; simply the standard estimate of the risk difference (we discuss deriving our measures of uncertainty below which account for the repeated measures structure of the data). For the next two estimators, we either plug in an estimate

of $Q$ based upon logistic regression or on tmle/SL estimate into (3.7). The gold-standard estimator (less biased, asymptotically normally distributed) is that based upon targeted maximum likelihood estimation (tmle), but we estimate the others to provide some context for interpretation.

Whereas the unadjusted and standard regression estimators are easy to motivate and understand, they are based on faulty assumptions that can result in bias in the resulting estimator. For unadjusted, it is the implicit unmeasured confounding assumption, for the standard regression, it is the assumption of a logit-linear model. The tmle has essentially no modeling assumptions. In the case of our estimators, it is based on a $\hat{Q}$ that is an augmented version of our original SL fit described above. Specifically, it is simply adding a covariate that has the property of accounting for any residual confounding remaining in the SL fit as well as "smoothing" the estimator so it has a predictable (normal) sampling distribution so that formal statistical inference is possible. Note, this would not be the case if a pure machine-learning approach was used (in our case, simply plugging in the original SL fit as the $\hat{Q}$). The details of the estimator can be found in [2]; we used the tmle package available in the programming language R [53, 56]. Note we do the estimates of the average treatment effect both stratified by clinic as described, but also pooled over clinics, or simply by adding one more outer, empirical average:

$$\Psi(P_n) = \frac{1}{n} \sum_t \sum_r \sum_i \sum_j I(R_{ij} = r) * I(T_{ij} = t) * (\hat{Q}(1, W_{ij}, t, r) - \hat{Q}(0, W_{ij}, t, r)), \quad (3.9)$$

where, $n = \sum_{i=1}^{m} n_i$, is the total number of observations in the data. The only difference of this estimator from a weighted average of the clinic-specific estimators is that we do the tmle-step over the entire data, thereby gaining some potential power in estimation. Again, we repeat this for 3 different plug-in estimates of $Q$ as discussed above. Note, that the actual sums are over only the observations with no missing values for all relevant variables; thus, the divisor in the averages, e.g., in (3.9), are adjusted accordingly.

### Heterogeneity of DIABETIMSS Impacts

To examine the heterogeneity of the intervention impacts across clinics, we simply estimate the average treatment effects stratified by clinic, which gives a directly accessible picture of between clinic heterogeneity. However, we also examine further sources of heterogeneity by using a combination of exploratory statistical methods (such as tree regression [57]) as applied to a transformation of the so-called blip-function. Ignoring clinic and year for this exposition (though extensions to repeat different years and clinics is trivial), consider simplified data $O = (W, A, Y)$. The blip function is defined as:

$$blip(W) \equiv E(Y \mid A = 1, W) - E(Y \mid A = 0, W) = Q(1, W) - Q(0, W). \quad (3.10)$$

Note, it is a measure of the DIABETIMSS impact on subjects with the same confounders, $W$, so in our case, we examine how the treatment impact can vary across the joint distribution

of confounders. We could simply plug in estimates of $Q$ as we did above, either using SL or TMLE. In this, we explored variation in the treatment effect by sub-populations by using tree regression on a transformed outcome

$$Y^* \equiv \hat{Q}(1, W) - \hat{Q}(0, W),$$

where $\hat{Q}(a, W)$ is an estimate of $E(Y \mid A = a, W)$. $Y^*$ is obtained in the following steps:

1. build a training dataset with indicator of glucose control as outcome, and all the variables we used as predictors (minus clinic ID) as predictors,
2. build a Super Learner model with same algorithms as above("SL.mean", "SL.earth", "SL.xgboost", "SL.glm", "SL.glm.interaction") on the training dataset,
3. build 5 test datasets, one for each control clinic, and set the indicator of DIATBETIMSS program status (enrolled to DIABETIMSS) as 1, keeping the $W_i(t)$ for the clinic fixed at the original values,
4. fit the Super Learner model on the test datasets, and compute the predicted outcome as $\hat{Q}^*(1, W)$ for each observation,
5. build 5 test datasets, one for each control clinic, and set the indicator of DIATBETIMSS program status (enrolled to DIABETIMSS) as 0,
6. fit the Super Learner model on the test datasets, and compute the predicted outcome as $\hat{Q}^*(0, W)$ for each observation,
7. compute $Y^* = \hat{Q}^*(1, W) - \hat{Q}^*(0, W)$ for each observation.

Because $E(Y^* \mid W) = blip(W)$, one can regress $Y^*$ against $W$ to explore what factors are most important for impacting the treatment effect. There are more formal ways of deriving so-called optimal treatment regimes and estimates of the future benefit of implementing such a rule [58]. However, there is little reason to think the program should decrease the probability of glucose control, and empirically this was born out (as discussed below in Results), so finding an optimal rule seems most likely trivial: enroll everyone in DIABETIMSS. One could also estimate the optimal subset of patients to enroll in DIABETIMSS if there was resource constraints (could not give the program to everyone), but that is beyond the current scope of our chapter. We simply used tree-regression, regressing $Y^*$ against $W$ to find sub-groups (defined in $W$-space) that have similar intervention impacts. We also simply examined the empirical distribution of the estimated blip function to see if the intervention impact seem to be uniform across patients, or was concentrated in a subset of patients. The combination of these descriptive approaches provides some indication of the potential benefit of targeting the program to subsets of patients, and to highlight subsets of patients that might not be currently benefiting from the program as much as desired.

We were also able to use models developed on the six clinics to predict the impact of the DIABETIMSS program on the clinics who have no enrolled participants (we refer to them as "control clinics"). This demonstrates how one can use the local patient characteristics, potentially clinic characteristics as well, to predict the impact of the program. This will be particularly important if there is large heterogeneity in the impact across clinics and subjects within clinics.

### Missing Data

We performed complete case analysis. Given at this point, there were no other (outcome) predictive covariates available to explain missingness beyond what we used in our models means that the conditional regression estimates (either from parametric models or SL) assume the data are missing at random (MAR), meaning that the outcome is independent of whether or not data were missing, *conditionally on the predictors in the model*. This avoids additional errors that can result from misspecification in either inverse weighting related models or imputation models. However, it does mean the marginal estimates are on the subset of the population that have non-missing data as we do not extrapolate to observations without missing data. Future work will involve sensitivity of results to other methods of handling missing data.

## Sensitivity analysis

We performed the same set of analyses, but also adjusting for the process of care indicators in addition the original adjustment variables described above. Because of the time-resolution available for this analysis, and that these indicators could be influenced by being enrolled in the DIABETIMSS program (e.g, comprehensive foot exam), we were worried about mixing up confounders and downstream causes of the program and potentially biasing our results. However, there is also some risk of not including them, as they might be argued to be important confounders. Thus, we repeated are analyses with these indicators included in adjustment set to see 1) if the overall associations were importantly different and 2) if we could provide evidence for the issues surrounding these indicators. Thus, in addition to duplicating the analyses, we also look at the distribution of the estimated propensity score (the $g(W)$ above). If there is strong relationship of being enrolled in the program and these indicators, then one would expect the propensity score distribution to shift to more extreme numbers (closer to 0 and 1) than the distribution of the estimate of the propensity score that does not contain those variables. If there is a significant proportion of the distribution close to $\hat{g}(W) = 0$, then estimation adjusting for these variables becomes problematic. Data at a more refined time-scale would be necessary to tease apart cause and effect if these indicators were thought to have a large influence on the outcome.

## Simulations for evaluating relative performance of estimators

To explore the relative merits of a more non-parametric (machine learning) approach relative to standard regression analyses, we conducted a set of simulations. We based the simulations closely upon the actual data, using a specific clinic's data to estimate the data-generating distributions. We used flexible, machine (SL) methods to estimate both the outcome and treatment models. We then ran simulations based upon this model (can be thought of as a semi-parametric bootstrap) and one where more non-linearity was entered into estimation of the prediction model (details to follow). We then compared the performance of the estimates and the confidence intervals of competing methods.

The purpose is to show the greater robustness of the Targeted Learning approach to estimation of adjusted associations.

### Details of the data-generating distribution for prediction, $Q$

We used clinic C at year 2016 to motivate the data-generating distribution. We could have chosen others, but we wanted sufficient number of observations (6,793) to get a reasonable estimate of the distribution of covariates, and both the treatment and prediction models.

We then fit the response variable, (the HbA1c control indicator), on the DIABETIMSS status and covariates (the restricted set that does not include quality of care indicators 1 to 7, as in the primary data analysis) and obtained the estimated $\hat{g}(W) \equiv \hat{P}(A = 1|W)$, and estimated $\hat{Q}(A, W) \equiv \hat{E}(Y|A, W)$. These were then treated as the true distributions in repeated simulations. We also did a separate set of simulations by augmenting the prediction model in the following way, making a $\hat{Q}(0, W)_{aug}$ and $\hat{Q}(1, W)_{aug}$:

$$logit(\hat{Q}(0, W)_{aug}) = logit(\hat{Q}(0, W)) - A - A \times logit(\hat{Q}(0, W)) - A \times logit(\hat{Q}(0, W))^2,$$

$$logit(\hat{Q}(1, W)_{aug}) = logit(\hat{Q}(1, W)) + A + A \times logit(\hat{Q}(1, W)) + A \times logit(\hat{Q}(1, W))^2.$$

For the second set of simulations, we simply replace the original fit by this one when generating the simulated data. The estimators used to generate estimates from simulated data do not change, whether we use the original fit. Thus, we get the relative performance in two different data-generating situations. Of course, this is a small set of possible simulations, but the purpose is to show the robustness of one method (that it works regardless of data-generating distribution) versus that of competing standard methods.

### Generating the truth

The performance is based upon knowing the true parameter in these simulations. Given the black-box nature of the methods used to make the data-generating distributions, there is no simply analytical way to derive the true average treatment effect (ATE). Thus, we did so by simulation. Given the true parameter is $E(Q(1, W) - Q(0, W))$, we can generate the true value of this by 1) taking a random sample with replacement of W of very large size (in our case, 1,000,000). For each W, we get its corresponding $Q(1, W)$, $Q(0, W)$ and take there

differences, and average over the randomly drawn W's. This was repeated both by defining $Q$ to be original and augmented estimates from the clinic C data.

**Simulation algorithm**

The simulations are thus just repeated the estimators we compared among repeated samples from the two distributions. For each simulated data set, we:

1. draw 6,793 samples from $W$ with replacement;
2. generate random binomial variables $A$ from the W samples using $gW$ (based on fit, $\hat{g}(W)$ from original data),
3. generate binomial $Y$ using $Q(A, W)$ either on the original fit ($\hat{Q}$ or augmented $Q$, $\hat{Q}_{aug}$,
4. estimate the ATE from three different substitution estimators based upon (bold is how these are referenced in figures):
    - unadjusted logistic regression (standard unadjusted analysis - will be biased if there is $A$ is not randomized so thus there is confounding),
    - adjusted logistic regression (standard regression approach, and will be biased if true $Q$ is not well-approximated by a linear model), and
    - targeted learning (tmle) based upon SL fits of the $g$ and $Q$ as done in the main analyses.
5. calculate the 95% CI's for each of the estimators,
6. store the three estimates of the ATE,
7. repeated the 4 steps above for 1,000 times,
8. compare the repeated estimates of each type to the true value to get the mean-squared error (MSE), and
9. compare the 1000 95% CI's to true ATE to get the true coverage probabilities of the three approaches.

## 3.3 Results

### Summary Statistics

Table 3.1 reports the data dictionary for this chapter and table 3.2 displays the overall association (across the years) of each of the predictor variables and indicator of diabetes control. First of all, there is a a significant ($p < 0.001$) unadjusted positive association with

the participation in the DIABETIMSS program and having a positive indicator of glucose control (36 versus 32%). Not surprisingly, the recent history of HbA1c is a strong predictor of current HbA1c status: 61% of subjects that had HbA1c $< 7\%$ in the previous year had control in the following year, where only 18% of subjects that had lack of control the previous year, achieved control the following year ( $p < 0.01$). There is a significant positive association of age and the HbA1c indicator: 35% of patients of $\leq 55$ years had glucose control, whereas 37% of patients over 71 years had control ($p < 0.001$). No anthropometric nor nutrition related variables are related to glucose control. Those that had multiple risk factors had unexpectedly similar glucose as those with no risk factors (28 versus 30%; $p = 0.406$). Interestingly, those that smoke have higher positive HbA1c indicators (35% for smokers versus 29 % for non-smokers). There is a significantly higher average number of complications related to diabetes in subjects with lack of glucose control ($p < 0.001$).

| Variables | Type | Description | Values |
|---|---|---|---|
| **Treatment Variable** | | | |
| diabetimss | binary | Patient referred to DIABETIMSS | 0) No 1) Yes |
| **Covariates** | | | |
| edad | continuous | age | |
| sexo | binary | sex | 1) Female 2) Male |
| tipo_pac | categorical | type of patient | 1) Insured 2) Spouse of insured 3) Child of insured 4) Parents insured 5) Retired |
| anttab | binary | smoking habit | 0) No 1) Yes |
| pesoini | continuous | Weight at the beginning of the year ($kg$) | |
| tallaini | continuous | Height at the beginning of the year ($m$) | |
| imcIni | continuous | BMI at the beginning of the year ($kg/m^2$) | |
| edoNutricioIni | categorical | Nutrition status at the beginning of the year | 1) Underweight 2) Normal weight 3) Overweight 4) Obesity |
| sobObes | binary | Overweight / Obesity | 0) No 1) Yes |
| tot_enfcrondiab | continuous | Total number of diabetic complications | |
| indic10_prev | binary | Indicator 10: Having HbA1C <7% in the last measurement; or in the absence of HBA1 test fasting glucose <= 130mg/dl in the last 3 measurement in previous year | 0) No 1) Yes |
| facriesg | binary | Patients with Risk Factors (smoking; hypertension; dyslipidemia) | 0) No 1) Yes |
| year | categorical | record year | 2012 to 2016 |
| **Process-of-care Variable** | | | |
| indic1 | binary | Indicator 1: Referral to the screening for dyslipidemia by measuring total cholesterol in patients without previous dyslipidemia | 0) No 1) Yes |
| indic2 | binary | Indicator 2: At least one measurement of HbA1C | 0) No 1) Yes |
| indic3 | binary | Indicator 3: Comprehensive foot evaluation | 0) No 1) Yes |
| indic4 | binary | Indicator 4: Referral to the ophthalmologist | 0) No 1) Yes |
| indic5 | binary | Indicator 5: At least one nutritional counseling provided by the nutrition service | 0) No 1) Yes |
| indic6 | binary | Indicator 6: Overweight and obese patients who received metformin unless contraindicated | 0) No 1) Yes |
| indic7 | binary | Indicator 7: Patients with hypertension receiving inhibitors of angiotensin converting enzyme (IACE) or angiotensin-receptor blocker unless contraindicated | 0) No 1) Yes |
| **Outcome Variable** | | | |
| indic10_curr | binary | Indicator 10: Having HbA1C <7% in the last measurement; or in the absence of HBA1 test fasting glucose <= 130mg/dl in the last 3 measurement in current year | 0) No 1) Yes |

Table 3.1: Variables analyzed in the chapter

| Variables | HbA1c >= 7% | HbA1c <7% | Missing | Adjusted p-value |
|---|---|---|---|---|
| Referred to DIABETIMSS, n (prop.) | | | | <0.001 |
| No | 63284 (0.50) | 31225 (0.24) | 33258 (0.26) | |
| Yes | 16254 (0.54) | 8940 (0.30) | 4797 (0.16) | |
| Missing | 65391 (0.21) | 23556 (0.08) | 224410 (0.72) | |
| Previous HbA1c results, n (prop.) | | | | <0.001 |
| HbA1c >= 7% | 69031 (0.60) | 15161 (0.13) | 30918 (0.27) | |
| HbA1c <7% | 12707 (0.26) | 19724 (0.41) | 15628 (0.33) | |
| Missing | 63191 (0.21) | 28836 (0.09) | 215919 (0.70) | |
| Age, n (prop.) | | | | <0.001 |
| [0,53) | 39172 (0.55) | 12795 (0.18) | 19186 (0.27) | |
| [53,62) | 37784 (0.53) | 14612 (0.21) | 18642 (0.26) | |
| [62,71) | 38587 (0.53) | 18023 (0.25) | 16378 (0.22) | |
| [71, 116] | 29386 (0.47) | 18291 (0.29) | 15395 (0.24) | |
| Missing | 0 (0.00) | 0 (0.00) | 192864 (1.00) | |
| Nutrition status at the beginning of the year, n (prop.) | | | | 0.646 |
| Underweight | 462 (0.44) | 190 (0.18) | 409 (0.39) | |
| Normal weight | 24399 (0.51) | 10454 (0.22) | 13326 (0.28) | |
| Overweight | 59249 (0.52) | 25584 (0.23) | 28164 (0.25) | |
| Obesity | 60609 (0.53) | 27360 (0.24) | 27228 (0.24) | |
| Missing | 210 (0.00) | 133 (0.00) | 193338 (1.00) | |
| Sex, n (prop.) | | | | 0.004 |
| Female | 86565 (0.52) | 38609 (0.23) | 40071 (0.24) | |
| Male | 58364 (0.52) | 25112 (0.22) | 29530 (0.26) | |
| Missing | 0 (0.00) | 0 (0.00) | 192864 (1.00) | |
| BMI at the beginning of the year ($kg/m^2$), n (prop.) | | | | 0.901 |
| [11.2, 26.0) | 36448 (0.51) | 15636 (0.22) | 19581 (0.27) | |
| [26.0, 28.9) | 36040 (0.53) | 15495 (0.23) | 16969 (0.25) | |
| [28.9, 32.4) | 36242 (0.53) | 15979 (0.23) | 16096 (0.24) | |

| Variables | HbA1c >= 7% | HbA1c <7% | Missing | Adjusted p-value |
|---|---|---|---|---|
| [32.4, 85.4] | 35989 (0.52) | 16478 (0.24) | 16481 (0.24) | |
| Missing | 210 (0.00) | 133 (0.00) | 193338 (1.00) | |
| Height at the beginning of the year (m), n (prop.) | | | | 0.003 |
| [1.30, 1.50) | 37877 (0.54) | 16438 (0.23) | 16342 (0.23) | |
| [1.50, 1.57) | 39267 (0.53) | 17393 (0.23) | 17526 (0.24) | |
| [1.57, 1.64) | 33231 (0.52) | 14773 (0.23) | 16363 (0.25) | |
| [1.64, 2.10] | 34344 (0.50) | 14984 (0.22) | 18896 (0.28) | |
| Missing | 210 (0.00) | 133 (0.00) | 193338 (1.00) | |
| Weight at the beginning of the year (kg), n (prop.) | | | | <0.001 |
| [30, 63) | 37150 (0.52) | 15744 (0.22) | 18099 (0.25) | |
| [63, 72) | 36771 (0.52) | 16188 (0.23) | 17106 (0.24) | |
| [72, 82) | 35912 (0.53) | 15694 (0.23) | 16594 (0.24) | |
| [82, 198] | 34886 (0.51) | 15962 (0.23) | 17328 (0.25) | |
| Missing | 210 (0.00) | 133 (0.00) | 193338 (1.00) | |
| Obesity, n (prop.) | | | | 0.247 |
| No | 24861 (0.50) | 10644 (0.22) | 13735 (0.28) | |
| Yes | 119858 (0.53) | 52944 (0.23) | 55392 (0.24) | |
| Missing | 210 (0.00) | 133 (0.00) | 193338 (1.00) | |
| Patients with Risk Factors (smoking, hypertension, dyslipidemia), n (prop.) | | | | 0.022 |
| No | 24358 (0.50) | 9576 (0.20) | 14853 (0.30) | |
| Yes | 120571 (0.53) | 54145 (0.24) | 54748 (0.24) | |
| Missing | 0 (0.00) | 0 (0.00) | 192864 (1.00) | |
| Smoking Habit, n (prop.) | | | | <0.001 |
| No | 141903 (0.52) | 62119 (0.23) | 68196 (0.25) | |
| Yes | 3026 (0.50) | 1602 (0.27) | 1405 (0.23) | |
| Missing | 0 (0.00) | 0 (0.00) | 192864 (1.00) | |
| Type of Patient, n (prop.) | | | | 0.027 |

| Variables | HbA1c >= 7% | HbA1c <7% | Missing | Adjusted p-value |
|---|---|---|---|---|
| Others | 74686 (0.53) | 31175 (0.22) | 36123 (0.25) | |
| Parents insured/Retired | 70243 (0.52) | 32546 (0.24) | 33478 (0.25) | |
| Missing | 0 (0.00) | 0 (0.00) | 192864 (1.00) | |
| Total number of diabetes complications, n (prop.) | | | | <0.001 |
| >1 | 21752 (0.57) | 7894 (0.21) | 8605 (0.22) | |
| 0 | 77522 (0.50) | 36812 (0.24) | 40760 (0.26) | |
| 1 | 45655 (0.54) | 19015 (0.22) | 20236 (0.24) | |

Table 3.2: Distribution of HbA1c indicator among predictors, pooled over years and clinics. The adjusted p-value is derived by fitting a generalized estimating equations (GEE) with all the predictors, adjusting for patient ID. Then we did analysis of Wald statistic with binomial model and logit link to obtain the p-value.

### Missing Data

Table 3.2 shows the relationship of missing data for each of the variables to the missingness of the outcome. One can see there is extensive missing data, particularly for the HbA1c indicator. Over half (62%) of observations had missing HbA1c indicators, so a very significant proportion of missing data that does limit the confidence by which one can extrapolate the results to the larger population. As we have stated above, we did complete case analyses, but such a large share of missing data, there is no silver bullet for unbiased extrapolation of the results based on the non-missing observations to the larger population.

## Estimated Impact of Program

The results of the analyses estimating the impact of the DIABETIMSS program are shown in figure 3.1 and table 3.3. Comparing first at the TMLE results across clinics and pooled ("All" clinics), there looks to be a fair amount of variability in the treatment impact; clinic E shows an estimated improvement in the HbA1c indicator of around 2%, whereas results for F suggests an 8% improvement. The overall (pooled) estimated suggests a 5% improvement in glucose control (last row of table 3.3). Comparing the unadjusted to the two adjusted estimates (standard regression and machine-learning adjusted TMLE) shows strong evidence of confounding by the measured factors. For most clinics (and the pooled estimate) the adjusted estimates are generally more significant than the unadjusted (estimates move away from the null), suggesting that the DIABETIMSS was assigned with higher probability to patients with higher risk of disease.

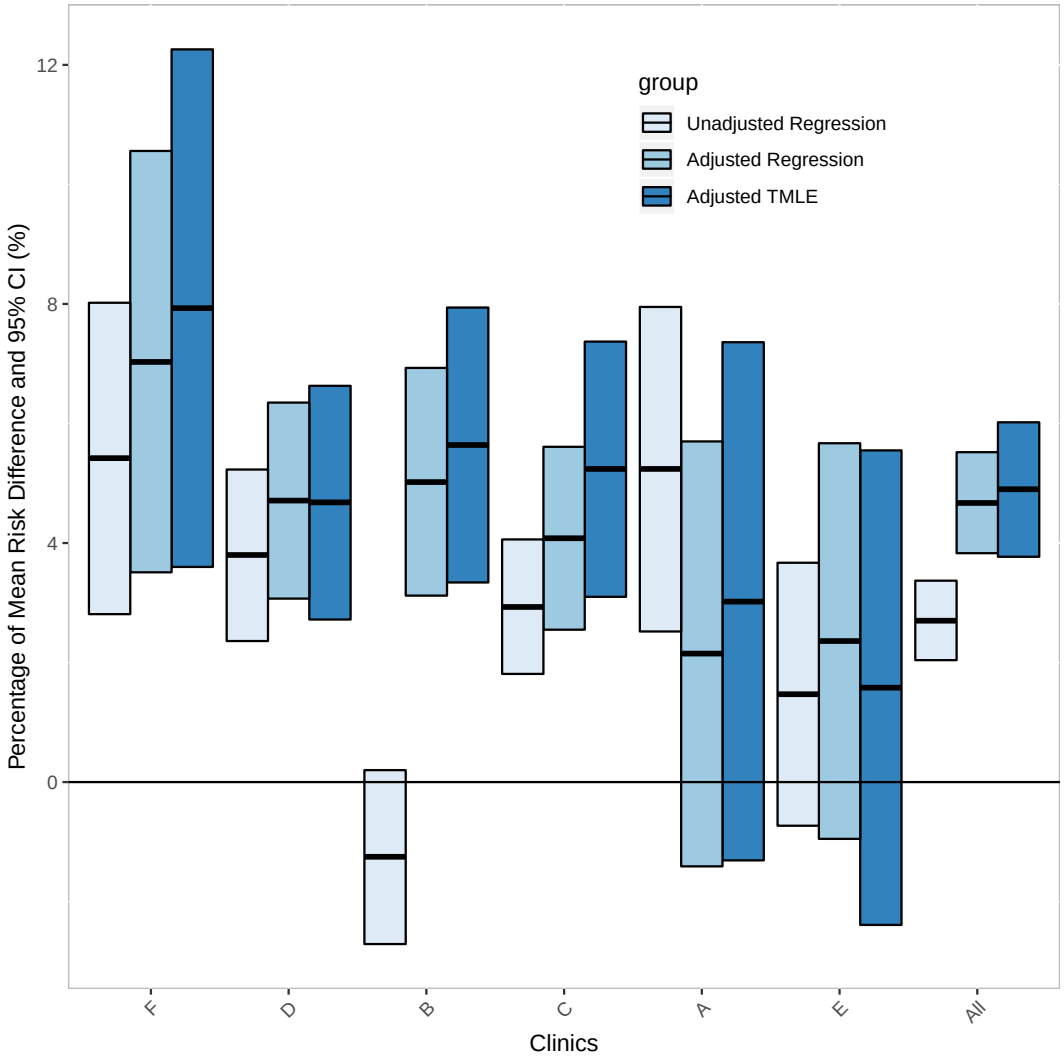


Figure 3.1: Targeted Learning adjusted associations of DIABETIMSS and glucose control for all DIABETIMSS clinics

## Exploration of heterogeneity among clinics

We explored potential heterogeneity across clinics in a few ways. First, we looked at the consistency of associations of predictor variables with the HbA1c indicator. Figure 3.2 shows the results of running logistic regression stratified by clinic and showing the resulting estimated log odds ratios by clinic. It shows, not surprisingly, a consistently strong assocation with the previous years HbA1c indicator, relatively consistent associations of DIABETIMSS across clinics (similar to the universal positive associaitons discussed above), and consistent associ-

| Clinic | DIABETIMSS | n | Unadjusted logistic regression | Adjusted logistic regression | TMLE |
|---|---|---|---|---|---|
| | | | % HbA1c <7% (95% CI) | % HbA1c <7% (95% CI) | % HbA1c <7% (95% CI) |
| | No | 4778 | 0.3599 (0.3517, 0.3680) | 0.3692 (0.3574, 0.3810) | 0.3694 (0.3577, 0.3812) |
| A | Yes | 573 | 0.4122 (0.3864, 0.4381) | 0.3907 (0.3564, 0.4249) | 0.3997 (0.3665, 0.4329) |
| | RD | | 0.0524 (0.0252, 0.0795) | 0.0215 (-0.0141, 0.0570) | 0.0302 (-0.0131, 0.0736) |
| | No | 6335 | 0.4027 (0.3950, 0.4103) | 0.4028 (0.3918, 0.4138) | 0.4034 (0.3926, 0.4143) |
| B | Yes | 2324 | 0.3901 (0.3777, 0.4025) | 0.4530 (0.4368, 0.4693) | 0.4598 (0.4444, 0.4752) |
| | RD | | -0.0125 (-0.0271, 0.0020) | 0.0502 (0.0312, 0.0693) | 0.0564 (0.0334, 0.0794) |
| | No | 11535 | 0.3419 (0.3369, 0.3470) | 0.3285 (0.3222, 0.3349) | 0.3291 (0.3228, 0.3354) |
| C | Yes | 3269 | 0.3713 (0.3612, 0.3814) | 0.3694 (0.3554, 0.3833) | 0.3815 (0.3673, 0.3957) |
| | RD | | 0.0293 (0.0181, 0.0406) | 0.0408 (0.0255, 0.0561) | 0.0524 (0.0310, 0.0737) |
| | No | 3500 | 0.3067 (0.2982, 0.3151) | 0.2911 (0.2801, 0.3022) | 0.2917 (0.2809, 0.3026) |
| D | Yes | 2208 | 0.3446 (0.3330, 0.3563) | 0.3382 (0.3252, 0.3513) | 0.3385 (0.3257, 0.3513) |
| | RD | | 0.0380 (0.0236, 0.0523) | 0.0471 (0.0307, 0.0635) | 0.0468 (0.0272, 0.0663) |
| | No | 1050 | 0.1570 (0.1483, 0.1657) | 0.1624 (0.1503, 0.1746) | 0.1623 (0.1501, 0.1744) |
| E | Yes | 229 | 0.1717 (0.1514, 0.1919) | 0.1860 (0.1551, 0.2169) | 0.1781 (0.1493, 0.2069) |
| | RD | | 0.0147 (-0.0073, 0.0367) | 0.0236 (-0.0095, 0.0567) | 0.0158 (-0.0239, 0.0555) |
| | No | 1160 | 0.2049 (0.1944, 0.2154) | 0.2370 (0.2219, 0.2520) | 0.2376 (0.2226, 0.2527) |
| F | Yes | 337 | 0.2590 (0.2352, 0.2828) | 0.3073 (0.2753, 0.3393) | 0.3169 (0.2844, 0.3494) |
| | RD | | 0.0542 (0.0281, 0.0802) | 0.0703 (0.0351, 0.1056) | 0.0793 (0.0360, 0.1226) |
| | No | 28325 | 0.3278 (0.3247, 0.3309) | 0.3225 (0.3184, 0.3266) | 0.3227 (0.3185, 0.3268) |
| All | Yes | 8940 | 0.3548 (0.3489, 0.3608) | 0.3692 (0.3617, 0.3768) | 0.3716 (0.3639, 0.3794) |
| | RD | | 0.0270 (0.0204, 0.0337) | 0.0467 (0.0383, 0.0552) | 0.0490 (0.0377, 0.0602) |

Table 3.3: Associations of DIABETIMSS program and the HbA1c indicator by clinic and pooled over all clinics. The first two rows in each clinic give the estimates of the proportion of subjects with HbA1c < 7% and 95% confidence intervals (CI). The last line is the risk difference (RD) as just the difference in these estimated proportions so it provides the measure of association of interest (estimated change in proportion of those with HbA1c < 7% in DIABETIMSS - those outside the program. We show three esitmators as discussed in text: unadjusted, adjusted within a linear-logistic regression and finally using targeted maximum likelihood estimation (TMLE).

ations with age. The other variables have less consistent associations, being positive in some clinics, negative in others. To explore whether some clinics had very different distributions of predictors, we performed a standard principle components analysis and colored the points on a resulting PCA plot by clinic (see Figure 3.3), which shows consistent overlap among clinics. This suggest there is no dramatic differences in covariate distributions among the 6 clinics. Finally, we examined the distribution of treatment impacts across all the individuals in the study. In figure 3.4, we plot the estimated $blip(W)$ function across index of individual after sorting by the magnitude of this blip function. In an idealized situation where all subjects had the same treatment impact (same blip function), then this plot would look like a horizontal line right at the average treatment impact. In practice, even if the treatment impact is homogeneous, given estimation error one would expect some departure from this

line. In our case, their appears to be a relatively notable departure from homogeneity, such that around 20000 units (individual/years) have impacts greater than the average impact, whereas the majority of subjects have impacts below this average impact. Very few subjects have estimated negative impacts, and this could be due to random estimation error, not actual negative effects of DIABETIMSS among the small proportion of subjects that have negative estimated blip functions (our $Y^*$ defined above in equation 3.10).

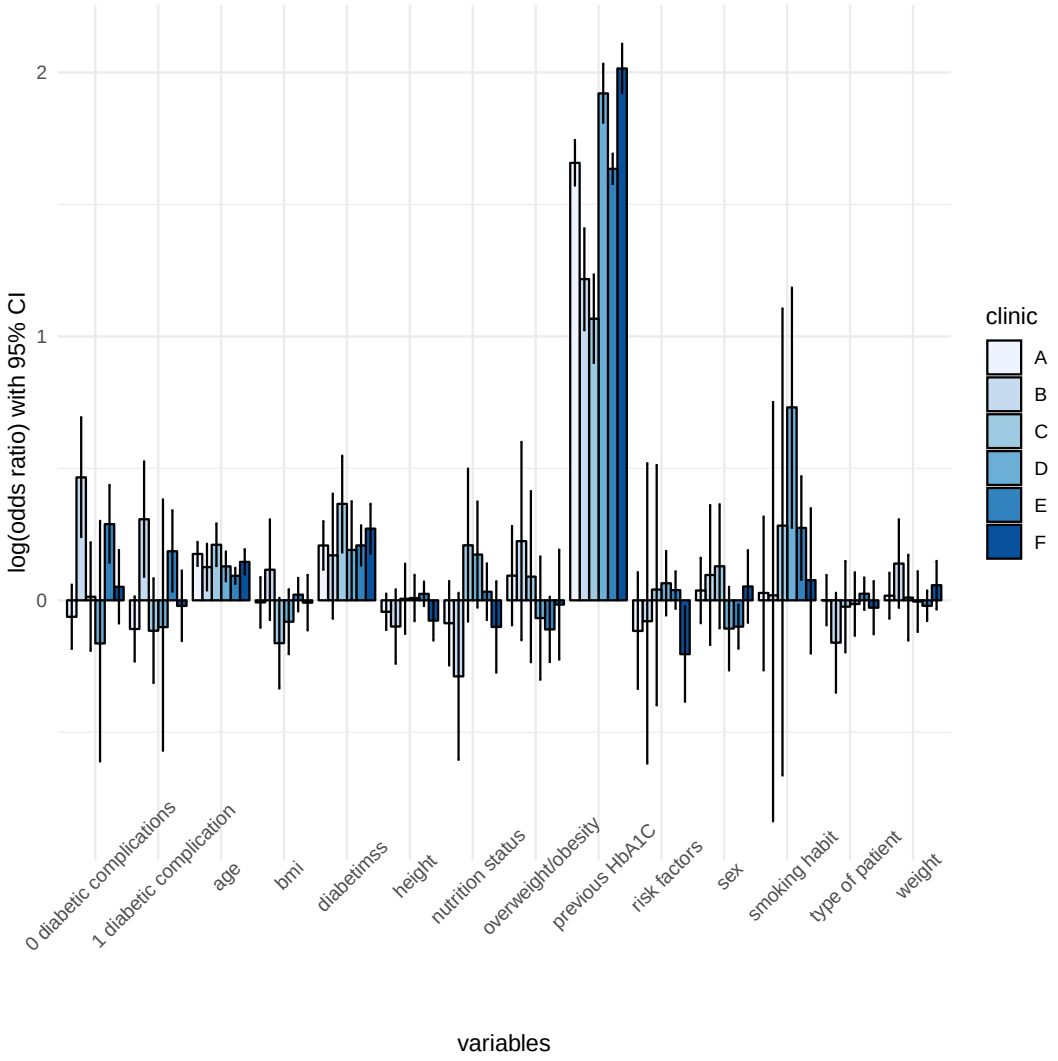


Figure 3.2: Comparison of Associations of Covariates and outcome by Clinic. The last 2 variables: tot_enfcrondiab0 and tot_enfcrondiab1 stands for 2 nominal levels of tot_enfcrondiab, which is total number of diabetes complications. The 3 levels of tot_enfcrondiab are: 0, 1, $> 1$

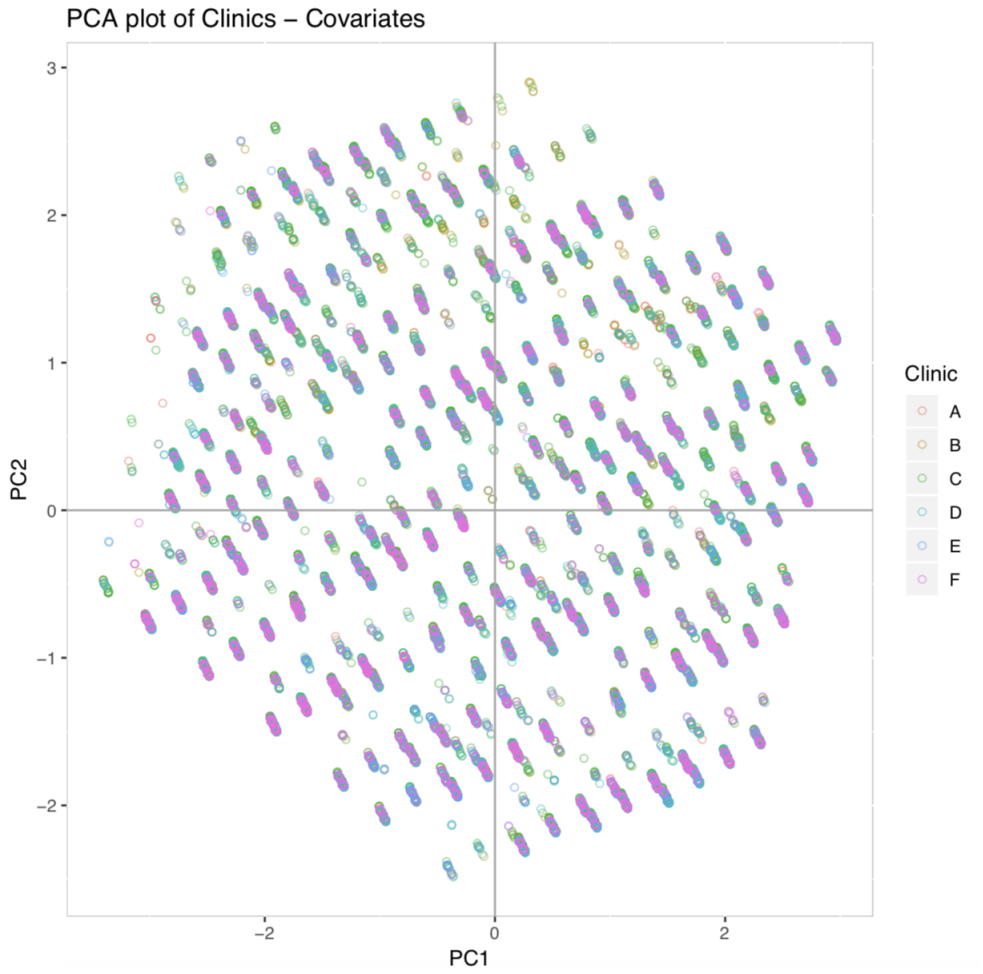


Figure 3.3: Principle components analysis of DIABETIMSS clinics. Note that clinic E does not appear distinct from other clinics.

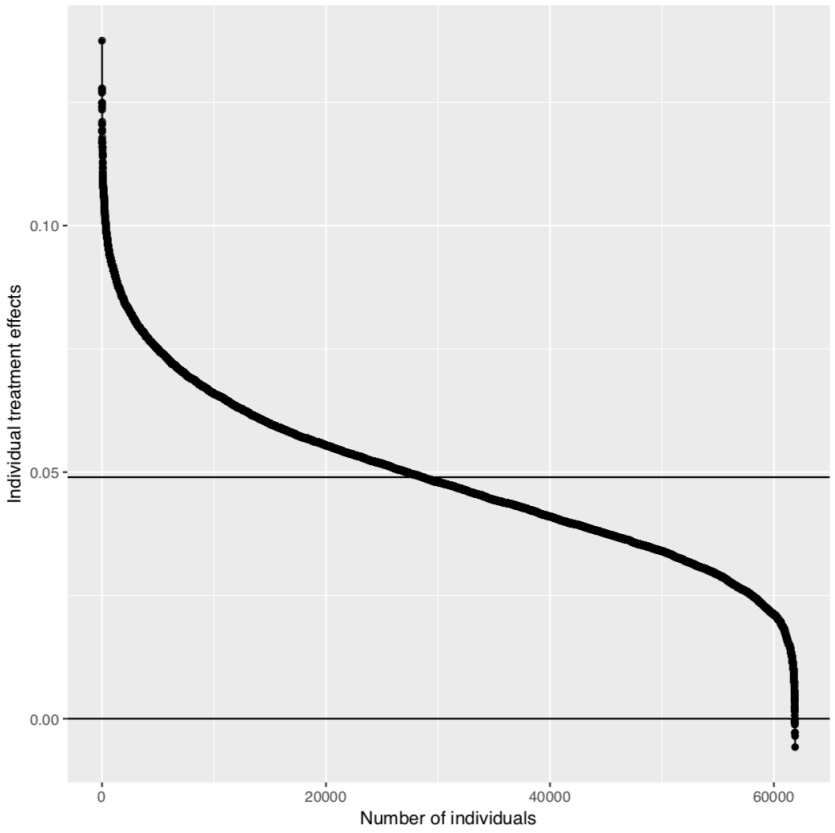


Figure 3.4: Distribution of DIABETIMSS treatment impacts among all subjects in DIABETIMSS clinics.

Finally, we attempted to explain the heterogeneity seen in figure 3.4 by performing tree regression [57] on the blip-function transformed data ($Y^*$) to explore the factors most responsible for differences in the treatment impact. The results of this are presented in figure 3.5. Tree regression is a simple form of histogram regression based on binary splits on covariates. It results in distinct nodes (representing sub-populations) that "best" characterize the variability seen in the outcome (in our case, the blip function). One can see the terminal nodes (the smallest subgroups) vary in their treatment impact from relatively small (2.6% in left most node) to modestly larger than the average treatment effect (6.4 %). If one exams the variables that define these splits, if there is a general message, it is that those with fewer existing complications of diabetes appear to have a greater benefit from the program than those with more progressed diabetes. This is not surprising as the magnitude of the reversal of the disease progression is larger and harder to achieve among this subset. However, one sees no obvious sub-populations where either the program is universally effective or visa versa. Thus, for any group, using the average impact estimated (around 5% improvement) is not a unrealistic estimate.

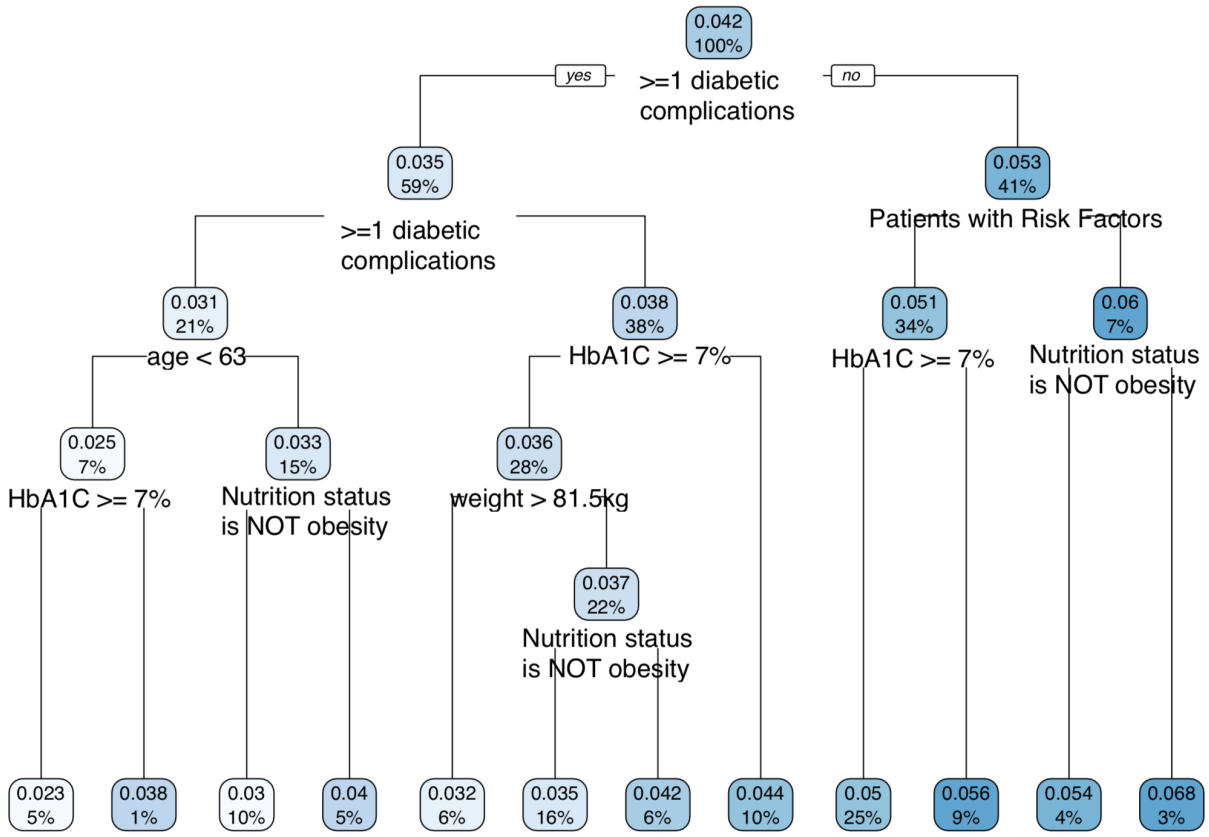


Figure 3.5: Tree diagram showing the distribution of treatment effects based on fit to data from all DIABETIMSS clinics and applied to all control clinics

## Predicting benefit of DIABETIMSS in control clinics

As described above, we also predicted the impact on a patient by patient basis for the control clinics, and show the results in figure 3.6. As one can see, the predicted impacts are quite similar to what was observed empirically in the DIABETIMSS clinics, that is, there is some variation but one would expect above a 5% improvement in HbA1c indicators (on average) if the program were implemented in these clinics.

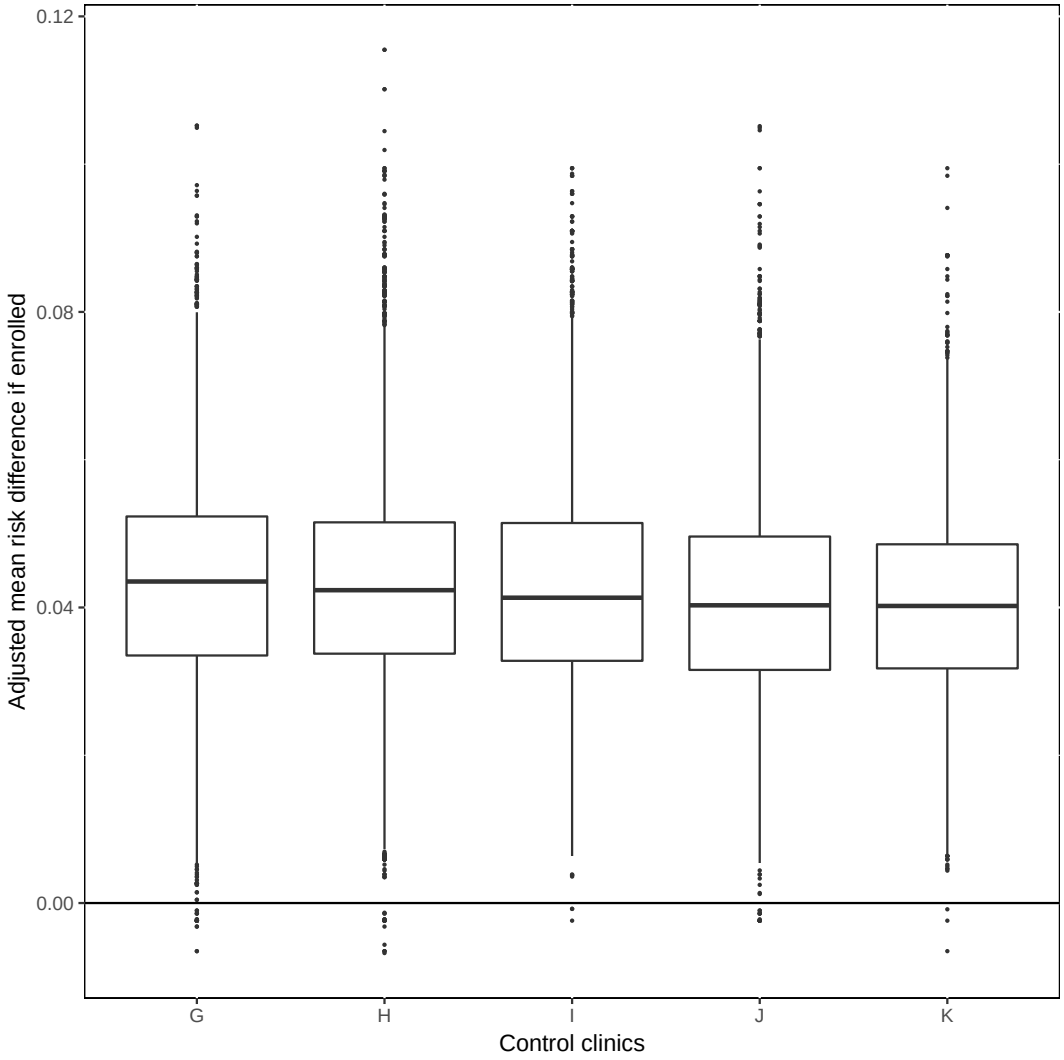


Figure 3.6: Boxplot (showing interquartile range) of predicted impact (blip function) of implementing DIABETIMSS program in control clinics, based upon TMLE fit of $Q$ in DIABETIMSS clinics

## Simulations

Figure 3.7 shows the plots of the sampling distribution along with the mean of the 3 estimators and and the true mean (black line). In addition, the caption contains specific numbers regarding the performance of the different estimators. The left and middle plots are the estimates of one component of the ATE (adjusted mean when $A = 0$ and $A = 1$, respectively). The farthest right is that of the parameter of interest, the ATE. One can see small reduction in bias in the tmle, versus the standard adjusted and unadjusted. However, even the mean

of the unadjusted estimates is close to the true value, and its confidence interval has nearly perfect 95% coverage, so there is very little room for improvement.

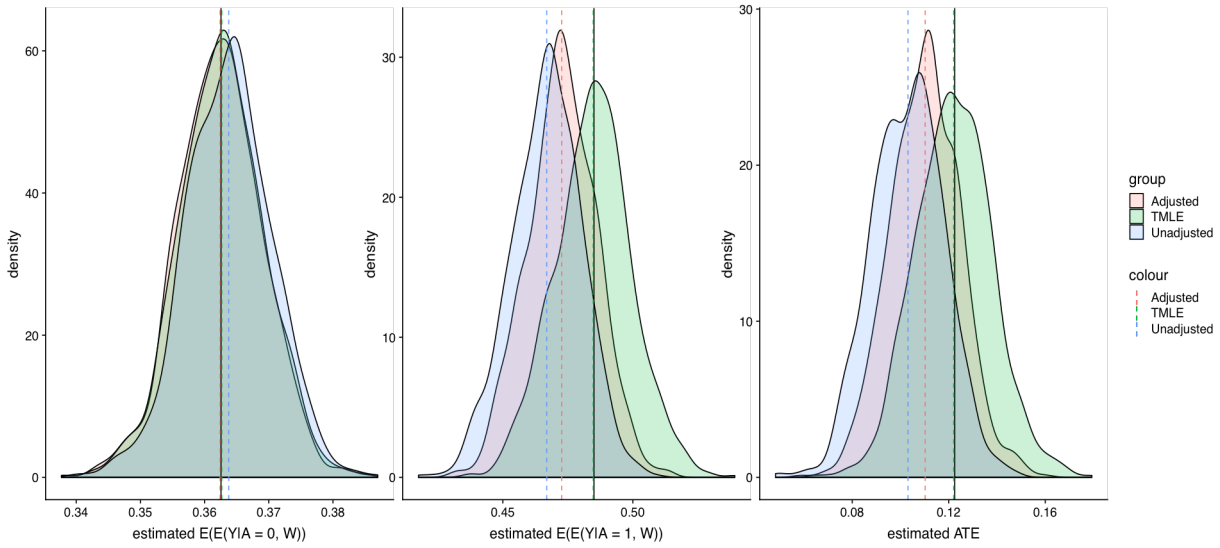


Figure 3.7: Distribution of model estimation using original data parameters. Dashed lines are the mean values. For Q0W: unadjusted (MSE = 4.56e-05, coverage = 94.0%); adjusted(MSE = 4.25e-05, coverage = 94.6%); TMLE(MSE = 4.34e-05, coverage = 94.8%). For Q1W: unadjusted (MSE = 0.0005, coverage = 72.5%); adjusted(MSE = 0.0003, coverage = 82.7%); TMLE(MSE = 0.0002, coverage = 92.7%). For ATE: unadjusted (MSE = 0.0006, coverage = 73.9%); adjusted(MSE = 0.0004, coverage = 86.5%); TMLE(MSE = 0.0002, coverage = 94.2%).

This is why we also used a augmented distribution to examine the relative performance when there is potential confounding by measured covariates as well as important non-linearities in the true prediction model. Figure 3.8 shows of the results of these simulations, along with detailed information on the relative importance. shows the three distributions with their asymptotic mean values, MSE and coverage rate for the 95% confidence interval. Clearly, the performance of the tmle estimator is far superior to the other two simpler estimators - they fail to pick up the confounding and are poor approximations for the true prediction model.

The message is that, if simpler, more parametric approaches work, so does tmle (simulation 1). However, tmle still works in cases where they fail (Simulation 2).

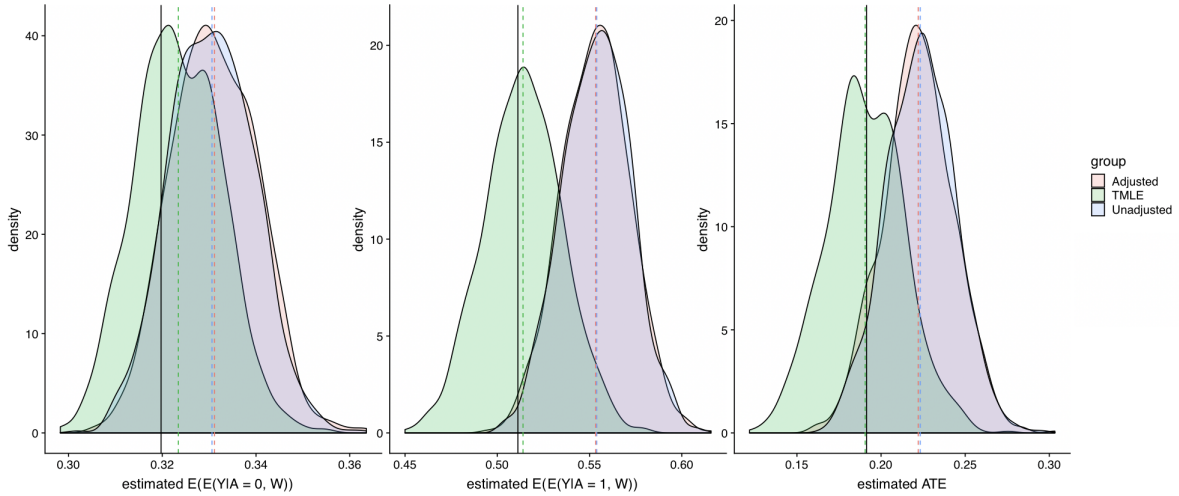


Figure 3.8: Distribution of model estimation using more variant data parameters. Dashed lines are the mean values. For Q0W: unadjusted (MSE = 1.97e-04, coverage = 77.1%); adjusted(MSE = 0.0021, coverage = 74.6%); TMLE(MSE = 0.0008, coverage = 94.2%). For Q1W: unadjusted (MSE = 2.15e-03, coverage = 33.5%); adjusted(MSE = 2.11e-03, coverage = 36.4%); TMLE(MSE = 4.44e-04, coverage = 90.7%). For ATE: unadjusted (MSE = 1.42e-03, coverage = 64.1%); adjusted(MSE = 1.38e-03, coverage = 67.8%); TMLE(MSE = 5.22e-04, coverage = 90.2%).

## Sensitivity: including process-of-care indicators as confounders

The resulting adjusted targeted learning estimates are shown in Figure 3.9, analogous to those shown in figure 3.1. One can see what appear to be more variable results (possibly more unstable) results, but the overall estimate average across all clinics does not change greatly. Thus, adjustment by these indicators does not change the main conclusions of the chapter. One can see that the distribution of propensity scores (see figure 3.10) has a larger proportion of the distribution at very low values (near 0) when the process-of-care indicators are included in the adjustment set. This indicates other variable importance results (not included by available upon request) that suggest weak association of these indicators with the outcome, but strong correlation with the program, again suggesting they are problematic as confounders for our outcome (HbA1c indicator).

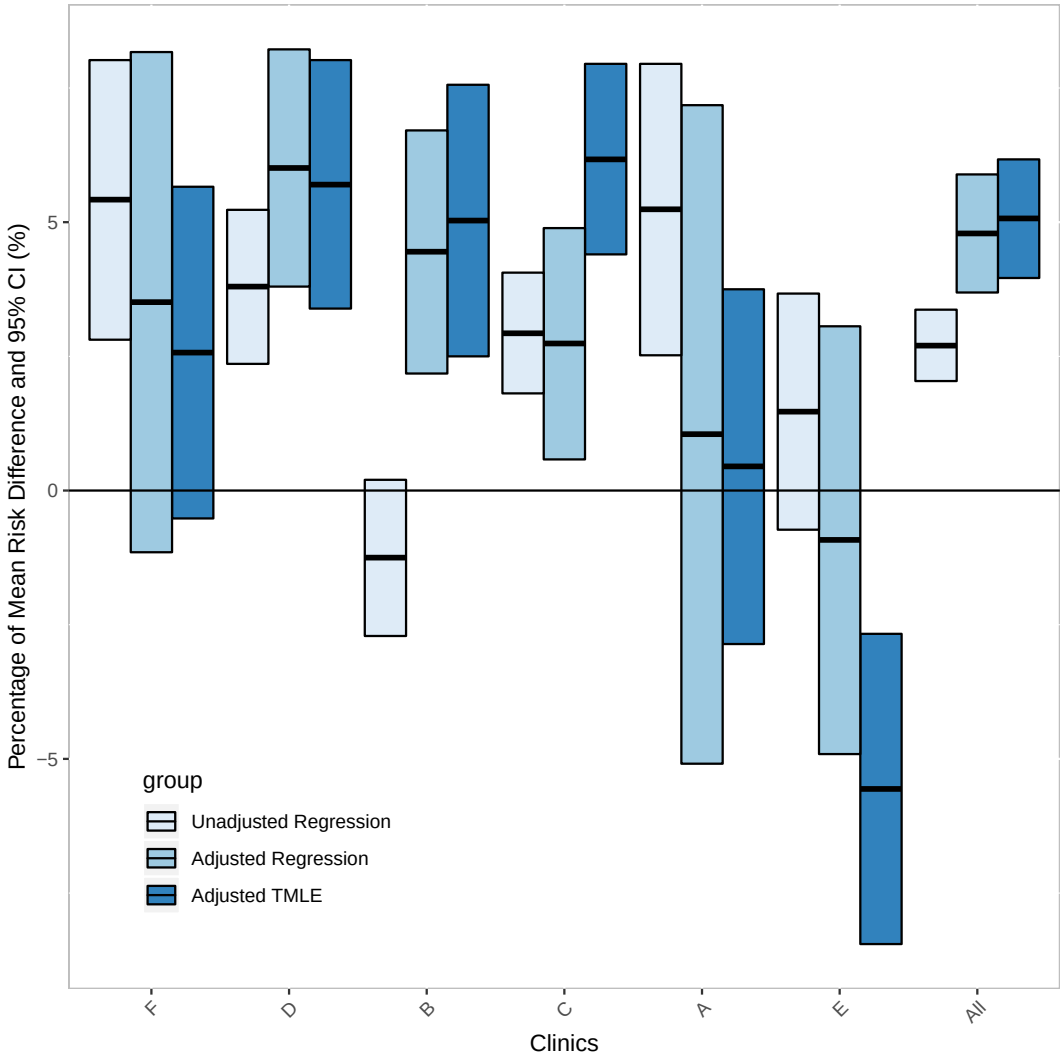


Figure 3.9: Targeted Learning adjusted associations of DIABETIMSS and glucose control for all DIABETIMSS clinics that includes adjust for process-of-care variables.

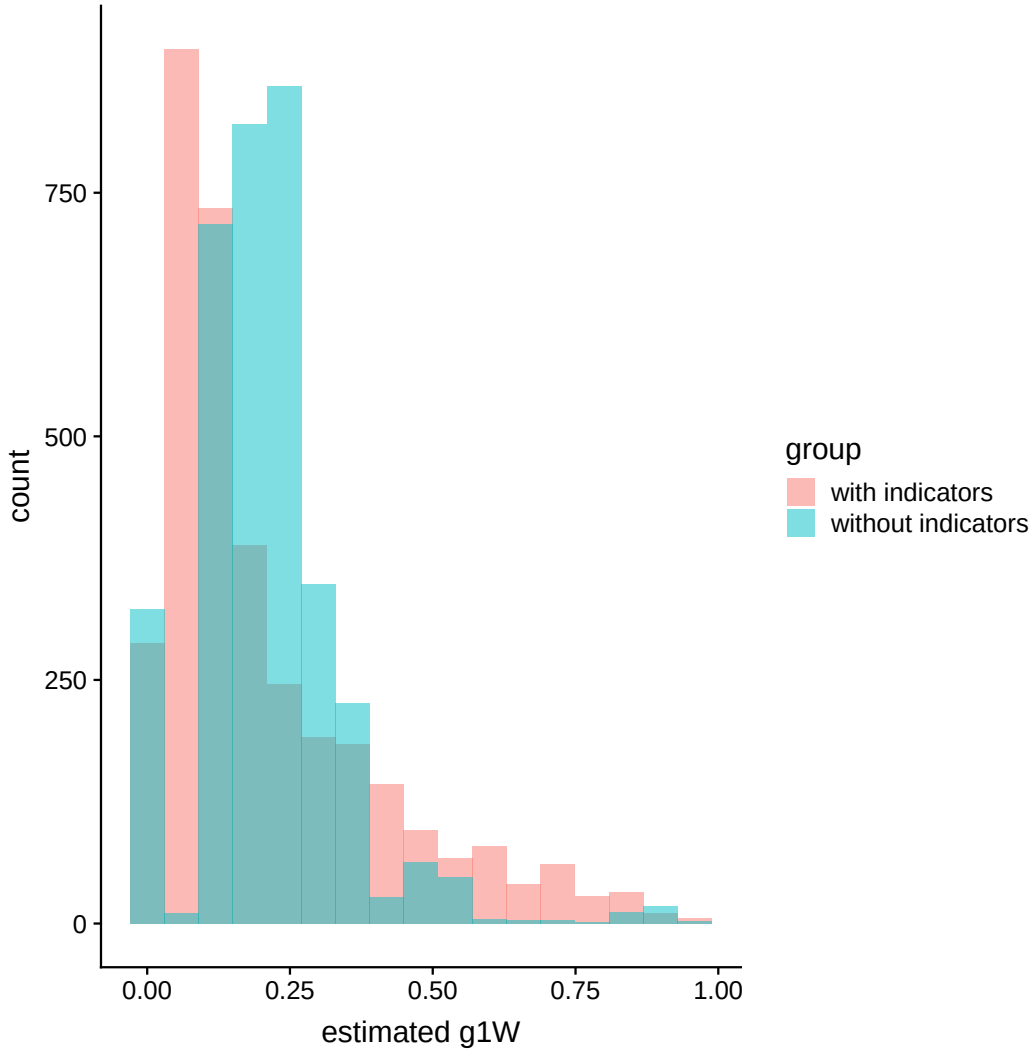


Figure 3.10: Distribution of estimated propensity scores, $g(W)$ both including and excluding the process-of-care indicators.

## 3.4 Discussion

The study provides evidence on the positive effect of DIABETIMSS program (pooled estimate of a 5% of improvement in glycemic control) and shows the potential and challenges in using routine observational patient data and machine learning methods to evaluate the performance of health interventions within complex health- care institutions to inform decision-makers.

DIABETIMSS was implemented to improve diabetes care and health outcomes by ad-

dressing three critical elements of the Chronic Care Model (CCM): 1) re-design of the delivery system through multidisciplinary teams, 2) decision support through evidence-based clinical guidelines, and 3) counseling and empowering of patients on self-management. Multiple clinical trials in different countries have tested these three elements, showing positive effects on the improvement of the processes of care and patients' outcomes [59, 60]. CCM has been increasingly advocated for effective management and control of NCDs within primary care [61]. Results from randomized controlled trials that have tested CCMs in primary care contexts in Europe show that compared to usual diabetes care, more patients reached treatment targets for blood pressure, and levels of blood sugar and cholesterol [62]. Experiences with CCMs in 8 Caribbean countries show improvements in baseline to follow up measures of blood glucose control and increases in the proportion of patients receiving a preventive practice or meeting quality-of-care indicators [63].DIABETIMSS evaluation results are consistent with other CCMs interventions, revealing a small but essential impact of this program with an overall pooled estimate of 5% improvement in glycemic control of T2D patients. Nonetheless, this slight increase in the percentage of T2D patients who achieved glycemic control call for further research, as IMSS' decision-makers require additional evidence to ascertain whether DIABETIMSS provide the interventions of the CCM optimally in compliance with evidence-based guidelines to assure high-quality care and better health outcomes [61, 64]. The evidence suggests that more significant benefits could be obtained through combining all six elements of the CCM that means incorporating the organizational changes that focus on creating a culture and mechanisms that promote safe, high-quality care, including the introduction of strategies to facilitate changes, and management of errors and quality control problems [60]. Another critical element of the CCM is the availability of timely and accurate health information systems to ensure program accountability and provide information for future improvement efforts [61].

The outcome variable of this study was HbA1C $< 7\%$. Since 2000, this goal is recommended by the IMSS diabetes clinical guidelines, independently of patient age. However, since 2016, American Diabetes Association (ADA), highlighted that HbA1C measurement may have limitations primarily in older adults who have medical conditions that increase red blood cell turnover (e.g., hemodialysis, recent blood loss or transfusion, or erythropoietin therapy), which can falsely increase or decrease A1C. Therefore, for adults $\geq$ 65years of age ADA recommends specific glycemic control goals of HbA1C $< 7.5\%$ for healthy older adults with few coexisting chronic illnesses and HbA1C $< 8.0\%$ or $< 8.5\%$ for older adults with multiple coexisting chronic illnesses or instrumental impairments or cognitive impairment [65]. If we apply the ADA recommendation to our study, this could probably increase the effect of the DIABETIMSS on glycemic control of older patients; yet, further analysis is recommended to support this hypothesis.

To date, diabetes research that used machine learning methods, was focused primarily on biomarker identification, prediction of diagnosis and diabetes complications, with low emphasis on evaluation of healthcare programs [66]. Our study is one of the pioneers to evaluate the performance of an ongoing health program using machine learning methods and routine observational patient data to inform decision-makers. The study showed both the potential

and challenges in using detailed observational patient data to evaluate the performance of a healthcare program. Though the estimates from standard regression were not radically different from those based upon less biased, machine learning methods, they do show enough different to be important, particularly when the impacts apply to so many potential patients. In addition, simulations based upon the data suggest the relative merits of using a more agnostic, machine-learning approach when they give different results than standard methods. In addition, the simulations show that the more complex targeted learning estimator does not harm performance when a simpler model provides an adequate approximation. Of course, one never knows whether or not standard methods will suffice at the beginning of a study, so standard methods are used, one is betting on the relationships being having a particular form. Thus, it increases the risk of misleading conclusions, with very little of any benefit relative to using more data-adaptive methods. The study not only shows the merits of using targeted learning approaches to evaluate the average performance of the intervention, but also to explore the heterogeneity of this performance across different clinics. Based upon the distribution of patient characteristics, the analyses also provide information regarding which clinics are most likely to benefit from future implementation of DIABETIMSS. The information provided could be the basis of informed cost-benefit analyses of DIABETIMSS or other programs.

The limits of the analysis are related to the limits of the available data. First, the data had significant missing values, particularly for the outcome, making extrapolation of the results on those non-missing observations more problematic for the entire population. In addition, more detailed measurements made a finer time scale would allow more refined estimation of the impact of DIABETIMSS and through which pathways the program operates to impact patient health. Given the limits of the resolution, however, the analyses were able to provide actionable information about the benefits of the program and how it might be optimally distributed within the relevant population of clinics. Beyond the specific application to DIABETIMSS, the combination of methods and data suggest this type of study is valuable for evaluating programs and treatments within large health care systems.

# Chapter 4

# Application of targeted learning in variable importance measure

## 4.1 Introduction

With overwhelmingly complex high-dimensional Electronic Health Record data nowadays, it's our primary goal in evidence-based medicine to establish the importance of numerous measured variables with regard to certain health outcomes, and hence assist researchers and healthcare workers to make optimal care decisions based on the entirety of the participants' data. There is not a universal definition of "most important" variables. One type of definition is motivated by prediction power, where the "most important" variables are those which accurately predict the outcome (like those generated from random forest [6]), and another type by causal association, where the "most important" variables are those whose change cause most changes in outcomes (like the average treatment effect of binary treatments [1]). Here, we focus on the latter one and define the variable importance as the amount of attribution of that variable towards changes in the mean outcome. There are plenty of medical literature measuring variable importance and building associations between these variables and outcomes of interest. Yet most of them only applies parametric models or evaluate the importance of the variable as how well the variable predicts the outcome [67]. In comparison, our method makes no parametric modeling assumptions and relies on data-adaptive ensemble machine learning models to estimate the data-generating distribution. The method we used in this chapter is developed based on the Non-Parametric Variable Importance (NPVI) estimator first proposed by Chambaz, Neuvial and van der Laan [68], which utilized the combination of machine learning and causal inference via targeted learning [2], to determine how changing the variable changes the outcome. There are several other approaches to measure the variable importance under different definitions utilizing targeted learning, including the collaborative targeted maximum likelihood estimation by Gruber and van der Laan [69, 70, 71], the data-adaptive target parameter by Hubbard, Kennedy and van der Laan [72], variable importance measure in longitudinal data by Diaz et al [73], variable importance

measure by Hubbard et al [74], the semiparametric regression model approach by Tuglus and van der Laan [75] and by Wang, Rose and van der Laan [76, 77]. Our approach has the following benefits. First, our approach does not dependent on arbitrary parametric assumptions between the covariates, the exposure and the outcomes (for example, coefficients in an arbitrary linear model). Second, our estimator is asymptotically linear (locally efficient) for which robust asymptotic inference is available. The relationship with causal intervention parameters also makes our estimator interpretable to health care workers. Third, our approach allows for variable importance comparisons that are comparable regardless of the original scale of the variable. We applied both the traditional linear model coefficients and our proposed estimator to measure the variable importance of mother's midlife eating behavior on child's BMI, adjusted for mother's adulthood stress, early-year socioeconomic status and health conditions, and child's age and sex, for participants involved in the National Heart, Lung, and Blood Institute Growth and Health Study [78]. We estimated the importance of five variables measuring mother's midlife eating behaviors, adjusted for eight variables measuring mother's early-year socioeconomic and health status, one variable measuring mother's adulthood strain count, and child's age and sex. Our results showed that if the mother has a higher level of drive for thinness, body dissatisfaction, bulimia, or interoceptive awareness, her child tends to have a higher level of BMI.

The remainder of the chapter is organized as follows. In section 4.2, we illustrated the methodology, including source of data, parameter of interest, and model specifications. In section 4.3, We presented the summary statistics and results. In section 4.4, we provided simulations for four estimators under linear and non-linear model settings, and evaluated their performance. In section 4.5, we discussed our findings.

## 4.2 Methodology

In this section, we first stated the source of data (section 4.2), then established the causal framework for our analysis(section 4.2), as well as the parameter of interest (section 4.2), and then we provided the specifications for the four estimators analyzed in this study (section 4.2).

### Data

The data from the National Heart, Lung, and Blood Institute Growth and Health Study (NGHS) [78]. Women from the original Richmond, CA site (N=883) were recruited for a follow-up study that extended the previous cardiometabolic aims to midlife. Comprehensive anthropometric, health, behavioral, psychosocial, and demographic data, collected annually from the original study period (1987-1997) when the women were age 9-10 to age 19-21, were combined with similar data collected when the women were age 37-43. To be eligible for the follow-up, women could not be pregnant, have given birth/miscarried within the last three months, or be incarcerated at recruitment. In this study, we used a sample of

women and children who completed the baseline questionnaire and the clinic visit (either at home, at Berkeley or a package with a scale and tape measure was mailed to them). Women and their biological children who either had a home visit or came to Berkeley were weighed and measured by trained staff. Women who completed the protocol via distance, where guided over the phone by a trained staff member to weigh and measure themselves and their children. Of the 883 original participants, 624 enrolled the follow-up study and completed the baseline questionnaire. Of the 624 participants 373 had 586 children who had measured weight, height and waist circumference. This study was approved by the University of California, Berkeley Institutional Review Board.

## Causal Framework

The variables we used in this study has of three components:

1. Exposures($X$): mothers' eating behaviors collected in the follow-up study (age 37-43): Four measures from the Eating Disorder Inventory that reflect dimensions of disordered thoughts, behaviors and attitudes toward eating, weight, body parts and emotions [79]:

    - drive for thinness (dt): The "drive for thinness" construct has been described as one of the cardinal features of eating disorders and has been considered an essential criterion for a diagnosis according to many classification schemes. The 7 items on this scale assess an extreme desire to be thinner, concern with dieting, preoccupation with weight and an intense fear of weight gain. Prospective studies have indicated that the drive for thinness scale is a good predictor of binge-eating and the development of formal eating disorders.
    - body dissatisfaction (bd): The body dissatisfaction scale consists of 10 items that assess discontentment with the overall shape and with the size of those regions of the body of extraordinary concern to those with eating disorders (i.e., stomach, hips, thighs, buttocks). One item on body dissatisfaction scale measures the feeling of bloating after eating a normal meal, a common feature of those who are dissatisfied with their body weight. Given the fact that body dissatisfaction is endemic to young women in Western culture, it is does not cause disorder alone; however, it may considered a major risk factor responsible for initiating and then sustaining extreme weight controlling behaviors seen in eating disorders.
    - bulimia (bul): The bulimia construct assesses the tendency to think about and to engage in bouts of uncontrollable overeating (binge-eating). The 8 items on this scale assess concerns about overeating and eating in response to being upset. The presence of binge eating is one of the defining features of bulimia nervosa and differentiates the bingeing/purging and restrictor subtypes of anorexia nervosa. Research has shown that binge eating is common in individuals who do not meet all of the criteria to qualify for a formal diagnosis of an eating disorder; however, in most cases, severe binge eating is associated with marked psychological distress.

- interoceptive awareness (int): The interoceptive awareness (int) scale consists of 9 items that measure confusion related to accurately recognizing and responding to emotional states. There is a "fear of affect" item cluster indicating distress when emotions are too strong or out of control that contrasts with an "affective confusion" item cluster indicating difficulty in accurately recognize emotional states. Confusion and mistrust related to affective and bodily functioning have been repeatedly described as an important characteristic of those who develop eating disorders.

An additional measure that captured non pathological tendency to overeat in the presence of food, reward-based eating drive (red), was also collected [80].

2. Covariates($W$): mothers' adulthood cumulative strain count collected in the follow-up study, her background information collected in the original study (age, race, parent education, household income, single or two parent household status, number of siblings, birth order, and her baseline BMI measure at age 10), and child's age and sex.

3. Outcome($Y$): child's BMI.

There exist a clear timeline for the three components, namely the covariates are determined before the exposures, and the outcome is measured at the same time with exposures. Therefore, we can build a Directed Acyclic Graph (DAG) [81] showing the causal relationship of the three components in Figure 4.1. Our hypothesis is that under mother's early year socioeconomic status, baseline BMI at age 10 and adulthood stress, and adjusting for child's age and sex, all the five exposures that measure mother's midlife eating behaviors will have a positive effect on child's BMI.

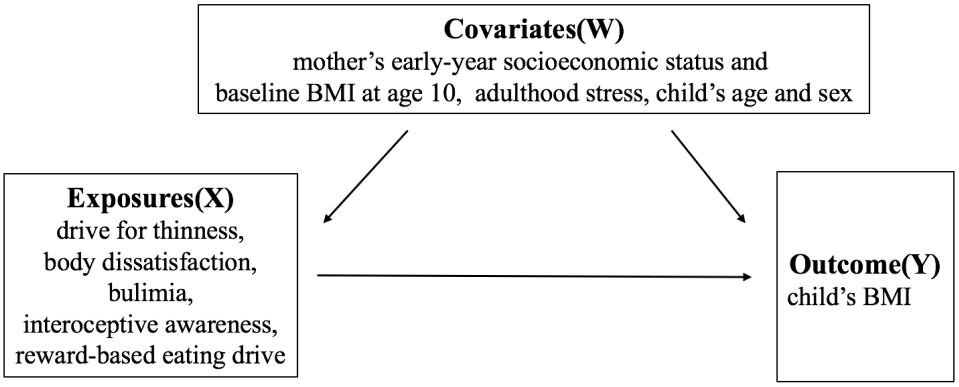


Figure 4.1: Directed Acyclic Graph (DAG) for the covariates, exposures, and outcomes

## Parameter of Interest

First, we build the causal framework for data analysis. Suppose the observed data structure is $O = (W, X, Y)$, where the $W$ represents the covariates, $X$ represents the exposures, and $Y$ represents the outcome. We wish to investigate the relationship between $X$ on $Y$, accounting for $W$. Taking $W$ into account is desirable because we know (or cannot rule out the possibility) that it contains confounding factors, i.e., common factors upon which the exposure $X$ and the response $Y$ may simultaneously depend. One classical approach to study such relationship is the Average Treatment Effect (ATE) [1], which investigates the causal effect of binary exposure $X$ on outcome $Y$. However, most variables measured in empirical studies are not binary, but feature a reference level $x_0$ with positive mass ($P(X = x_0) > 0$) and a continuum of other levels. Thus, we desire to upgrade the simple ATE parameter to one that measures the causal effect of continuous exposures on outcomes, and hence evaluate the relative importance of several variables (treated as exposures) with regard to outcomes. Such a parameter is named Targeted Maximum Likelihood Estimation-Non-Parametric Variable Importance (tmle.npvi) parameter [68]. In this study, we use the identity link for the $f$ function and reference level $x_0 = 0$ in the original definition of the tmle.npvi parameter. Our parameter of interest $\Psi$ is defined in proposition 1.

**Proposition 1.** *For all $P \in \mathbb{M}$:*

$$\Psi(P) = \frac{E_P\{X[E_P(Y|X,W) - E_P(Y|X=0,W)]\}}{E_P X^2}$$

*Proof.* The target parameter is formally defined as

$$\Psi(P) = \arg\min_{\beta\in\mathbb{R}} E_P[E_P(Y|X,W) - E_P(Y|0,W) - \beta X]^2$$

Thus, we have that

$$\Psi(P) = \frac{E_P\{X[E_P(Y|X,W) - E_P(Y|X=0,W)]\}}{E_P X^2}$$

The formal proof of proposition 2 can be found in [68] Section A.2. □

If the linear model is true, the implied causal parameter $\Psi(P)$ is $E_P(Y - Y_0)$, the average change in outcome if every observation were changed to have $X = 0$, where $Y$ is the observed outcome and $Y_0$ is the counterfactual outcome if $X = 0$. Furthermore, if $X = 0$ represents the a priori, lowest risk level, then $\Psi(P)$ can be interpreted as the attributable difference on $Y$ (risk if $Y$ is binary). We used the target parameter $\Psi(P)$ rather than $E_P[E_P(Y|X,W) - E_P(Y|X=0,W)]$ directly because targeting $\Psi(P)$ is more efficient as it is the projection on a linear model which allows estimators to borrow from the whole dose-response curve much like fitting a line $E_P(Y|X) = \alpha + \beta X$, where $\hat{\alpha}$ is an estimate of $E_P(Y|X=0)$, so we could just use the data at $X = 0$, $\sum_i Y_i \mathbb{1}(X_i = 0)/\sum_i \mathbb{1}(X_i = 0)$, or a line which borrows across the distribution of $X$.

The efficient influence function for our proposed parameter $\Psi(P)$ is defined in proposition 2.

**Proposition 2.**

$$\begin{aligned} D^*(P) &= \frac{1}{E_p(X^2)} X[E_P(Y|X,W) - E_P(Y|0,W) - X\Psi(P)] + \\ &\quad \frac{1}{E_p(X^2)}(Y - E_p(Y|X,W))(X - \frac{E_p(X|W)\mathbb{1}\{X=0\}}{P(X=0|W)}) \end{aligned}$$

*Proof.* The formal proof of proposition 2 can be found in [68] Section A.2. □

Here we emphasize that we do not assume a semi-parametric model

$$Y = \beta X + \eta(W) + U,$$

with unspecified $\eta$ and $U$ such that $E_P(U|X,W) = 0$. Setting

$$R(P,\beta)(X,W) = E_P(Y|X,W)E_P(Y|0,W)\beta X$$

for all $(P,\beta) \in \mathbb{M} \times \mathbb{R}$, the latter semi-parametric model holds for $P \in \mathbb{M}$ if there exists a unique $\beta(P) \in \mathbb{R}$ such that $R(P,\beta(P)) = 0$. Note that $\beta$ is always the solution to the equation

$$\beta E_P(X^2) = E_P\{X[E_P(Y|X,W) - E_P(Y|0,W) - R(P,\beta)(X,W)]\}.$$

In particular, if the semi- parametric model holds for a certain $P \in \mathbb{M}$, then $\beta(P) = \Psi(P)$ by Proposition 1. Furthermore, If the relationship of $Y$ on $(X,W)$ is truly linear, the target parameter $\Psi(P)$ can be seen as the projection of the difference $E_p(Y|X,W) - E_p(Y|0,W)$ on to the vector space of $X$, which has exactly the same definition as the parametric coefficients $\beta$ reported in linear regressions. On the contrary, if the semi-parametric model does not hold for $P$, then it is not clear what $\beta(P)$ could even mean whereas $\Psi(P)$ is still a well-defined parameter worth estimating. Further discussions can be found in [68] Section 4.2.

## Estimators

We evaluated four estimators for the target parameter $\Psi(P)$.

1. **Linear model plug-in estimator & linear model coefficient $\Psi_{lm}(P_n)$:**
   For linear model plug-in estimator, $E_P(Y|X,W), E_P(Y|0,W)$ are estimated by linear regression of outcome $(Y)$ on one of the five exposures (one of $X$) plus all eleven covariates (all $W's$), then the estimates are plugged in. The linear model coefficient estimator is the coefficient of $X$ in the linear model of outcome $Y$ on one of the five exposures (one of $X$) plus all eleven covariates (all $W's$). These two estimators are the same.

2. **Super Learner plug-in estimator $\Psi_{SL}(P_n)$:**
   $E_P(Y|X,W), E_P(Y|0,W)$ are estimated by an ensemble learner called Super Learner [82] with user-supplied machine learning algorithms, then the estimates are plugged in.

3. **Linear model target maximum likelihood estimator(TMLE) $\Psi_{tmle.lm}(P_n)$:**
   we update $\Psi_{lm}(P_n)$ using the targeted learning techniques with its efficient influence function defined in Proposition 2.

4. **Super Learner target maximum likelihood estimator(TMLE) $\Psi_{tmle.npvi}(P_n)$ (NPVI estimator)**:
   we update $\Psi_{SL}(P_n)$ using the targeted learning techniques with its efficient influence function defined in Proposition 2.

## 4.3 Results

Summary characteristics of the covariates are reported in Table 4.1. There are 624 participants who enrolled the follow-up study and completed the baseline questionnaire. Of the 624 participants 373 (60%) had 586 children that are used in this study, the other 40% of the participants are not included due to the lack of information on whether they have children or the lack of their children's records. Among the 586 children, only a small fraction of covariates are missing (max of 4%). We imputed the continuous covariates with the mean, and categorical ones with the mode. Observations missing exposures or outcomes are not included in the analysis. We performed data analysis with two distinct estimators, linear model coefficient $\Psi_{lm}(P_n)$ and NPVI estimator $\Psi_{tmle.npvi}(P_n)$. The linear model coefficient $\Psi_{lm}(P_n)$ is reported for the exposures in each of the five linear models (with same covariates and same outcome). The $\Psi_{tmle.npvi}(P_n)$ estimator is constructed on an initial estimate of a Super Learner. The Super Learner is constructed with general additive models (R package 'gam'-version 1.20 [83]), general linear models (R core package 'stats'-version 3.6.2), piecewise linear spline regressions (R package 'polspline'-version 1.1.19 [84]), and random forest (R package 'randomForest'-version 4.6-14 [85]).

| **Mother's Characteristic** | **Sample size** | **Mean (SD) or number (%)** |
|---|---|---|
| Baseline age | 586 | 9.93 (0.56) |
| Race | 586 | |
| Black | | 296 (51%) |
| White | | 290 (49%) |
| Parents' income | 564 | |
| $0-10K$ | | 115 (20%) |
| $10K-19999$ | | 97 (17%) |
| $20K-39999$ | | 161 (29%) |
| $40K+$ | | 191 (34%) |
| Parents' education | 585 | |
| College Grad+ | | 185 (32%) |
| High School or Less | | 117 (20%) |
| Some College | | 283 (48%) |
| Parents' marital status | 586 | |
| One parent household | | 204 (35%) |
| Two parent household | | 382 (65%) |
| Birth order | 586 | 2 (1) |
| Number of kids in household | 586 | 1 (1) |
| Baseline BMI | 581 | 18.4 (3.5) |
| **Children's Characteristic** | **Sample size** | **Mean (SD) or number (%)** |
| Age | 586 | 9 (4) |
| Sex | 586 | |
| Boy | | 293 (50%) |
| Girl | | 293 (50%) |

Table 4.1: Participant characteristics of mothers and children in the NGHS cohort study (373 mothers with 586 children)

The results of the data analysis is reported in Table 4.2. Overall, the estimated effects are consistent with similar scales across different estimators. In particular, drive for thinness (dt), body dissatisfaction (bd) and bulimia (bul) show consistently significant positive effects on the outcome, for both estimators with similar magnitude (dt: 0.08(0.02,0.15) for linear estimator and 0.12(0.08, 0.15) for NPVI estimator, bd: 0.07(0.02, 0.12) for linear estimator and 0.06(0.04, 0.07) for NPVI estimator, bul: 0.18(0.06, 0.30) for linear estimator and 0.2(0.12, 0.34) for NPVI estimator). Reward-based eating drive (red) has no effect on the outcome for both estimators (0.63(-0.06, 1.32) for linear estimator and 0.09(-0.28, 0.46) for NPVI estimator). interoceptive awareness (int) is the only exposure where the linear estimator and NPVI estimator have different results- linear estimator of shows no effect (0.08(-0.03, 0.18)) and NPVI shows a significant positive effect (0.15(0.06, 0.24)). The discrepancy could

be due to the very different statistical methods used in the linear and NPVI estimators, where the former assumes a linear relationship between the outcome and the exposure plus covariates, and the latter is much more flexible with model assumptions by including several non-linear machine learning models in the base learner. Thus we believe that the NPVI estimator is able to capture both linear and non-linear effect of the exposure on the outcome, and hence more robust and reliable than the linear estimator.

| **Variable** | **Linear estimator(95% CI)** | **NPVI estimator(95% CI)** |
|---|---|---|
| bulimia | 0.18(0.06, 0.30) | 0.2(0.12, 0.34) |
| interoceptive awareness | 0.08(-0.03, 0.18) | 0.15(0.06, 0.24) |
| drive for thinness | 0.08(0.02,0.15) | 0.12(0.08, 0.15) |
| body dissatisfaction | 0.07(0.02, 0.12) | 0.06(0.04, 0.07) |
| reward-based eating drive | 0.63(-0.06, 1.32) | 0.09(-0.28, 0.46) |

Table 4.2: Variable importance estimates with 95% confidence interval for mothers' eating behaviors, sorted by descending order of the magnitude of the NPVI estimator mean value. Variable with CI covering zero effects appear last.

## 4.4 Simulations

In this section, we performed two simulations based on the observed data. In both settings, the initial data sets are comprised of the outcome ($Y_0$), one of the exposures ($X_0$), and all eleven covariates ($W's$). In the first simulation setting (Algorithm 3), we assume a linear relationship between outcome and exposure plus covariates. Figure 4.2 illustrated the distribution of the mean values for the four different estimators in 1,000 iterations under this settings, and the estimator performance is summarized in Table 4.3. In this setting, all of the four proposed estimators in section 4.2 are unbiased with proper coverage for 95% confidence intervals (Table 4.3). The MSE of the SL plug-in estimator $\Psi_{SL}(P_n)$ and NPVI estimator $\Psi_{tmle.npvi}(P_n)$ are larger than that for the linear estimators.

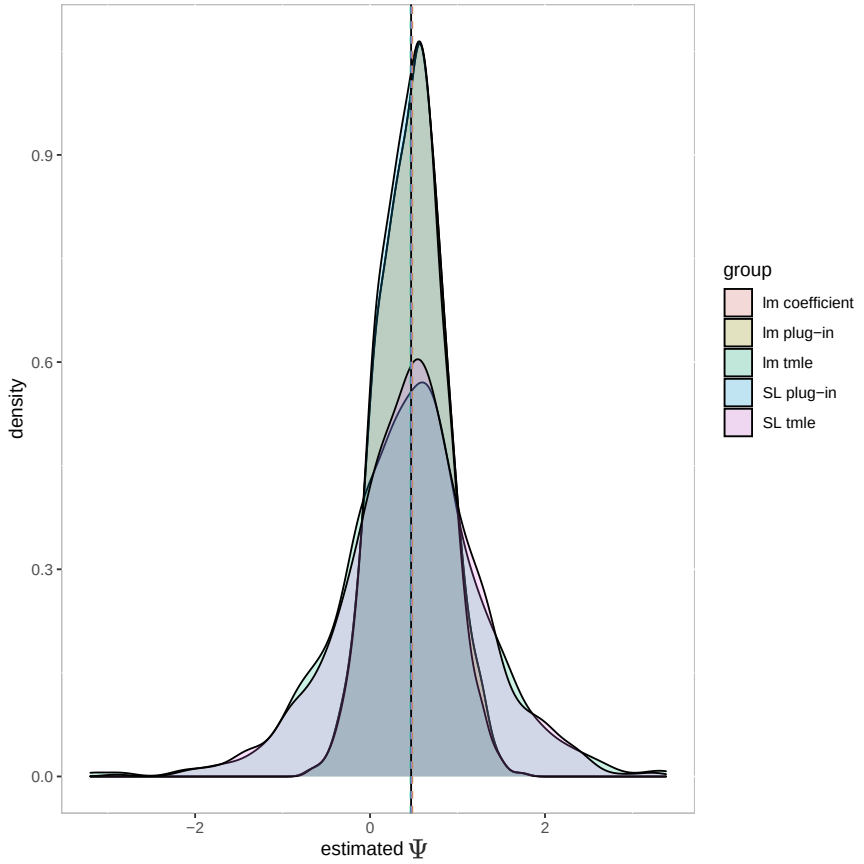


Figure 4.2: Distribution of the four estimator values under linear simulation settings. The black vertical line is the true value $\psi_0 = 0.47$, and the colored dashed lines are the mean values for each estimator.

| **Estimator** | **Mean** | **Bias** | **MSE** | **Coverage(%)** |
|---|---|---|---|---|
| lm coefficient & lm plug-in | 0.48 | 0.01 | 0.14 | 94.7 |
| lm tmle | 0.46 | -0.01 | 0.68 | 94.2 |
| SL plug-in | 0.47 | 0 | 0.13 | 95 |
| SL tmle | 0.47 | 0 | 0.61 | 93.9 |

Table 4.3: Performance of the four estimators under linear simulation settings. True $\psi_0 = 0.47$.

In the second simulation setting (Algorithm 4), we assume a non-linear relationship (combination between linear and random forest) between outcome and exposure plus covariates. Figure 4.3 illustrated the distribution of the mean values for the four different

estimators in 1,000 iterations under this settings, and the estimator performance is summarized in Table 4.4. In this setting, the linear estimator $\Psi_{lm}(P_n)$ is biased by 0.14, and its TMLE updated version ($\Psi_{tmle.lm}(P_n)$) reduced its bias dramatically to -0.03. The SL plug-in estimator $\Psi_{SL}(P_n)$ is biased by -0.08, and its TMLE updated version (NPVI estimator $\Psi_{tmle.npvi}(P_n)$) reduced its bias to -0.05. The TMLE updated estimators are not only less biased, but also achieved better coverage for the 95% confidence intervals (95.3% coverage for $\Psi_{tmle.lm}(P_n)$ and 94% coverage for $\Psi_{tmle.npvi}(P_n)$), whereas the plug-in estimators have lower coverage (88.9% coverage for $\Psi_{lm}(P_n)$ and 89.9% coverage for $\Psi_{SL}(P_n)$). The MSE of the four estimators are relatively small and of the same scale.

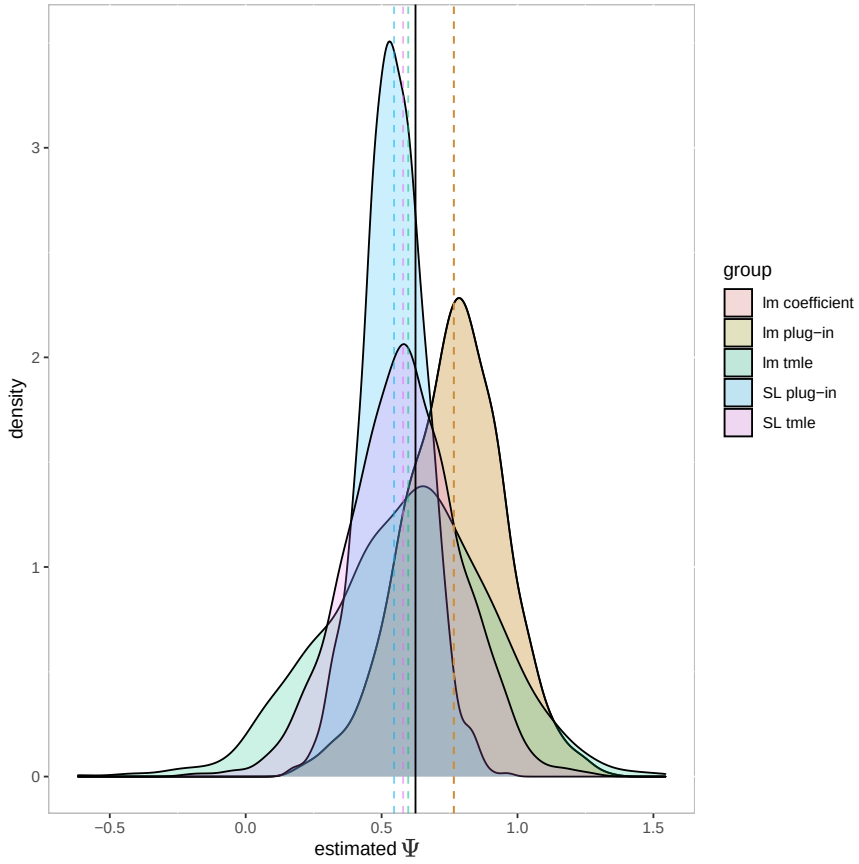


Figure 4.3: Distribution of the four estimator values under non-linear simulation settings. The black vertical line is the true value $\psi_0 = 0.62$, and the colored dashed lines are the mean values for each estimator.

| Estimator | Mean | Bias | MSE | Coverage(%) |
|---|---|---|---|---|
| lm coefficient & lm plug-in | 0.77 | 0.14 | 0.05 | 88.9 |
| lm tmle | 0.6 | -0.03 | 0.09 | 95.3 |
| SL plug-in | 0.55 | -0.08 | 0.02 | 89.9 |
| SL tmle | 0.58 | -0.05 | 0.04 | 94 |

Table 4.4: Performance of the four estimators under non-linear simulation settings. True $\psi_0 = 0.62$.

## 4.5 Discussion

The study provides evidence on the transmission effect of mother's midlife eating behaviors on child's BMI and shows the potential and challenges in using surveyed data and statistical learning methods to evaluate the influences of mother's health exposures on the next generation. The two distinct estimators (linear estimator and NPVI estimator) both showed positive effects of mother's drive for thinness, body dissatisfaction and bulimia on child's BMI, and both showed no effect of mother's reward-based eating drive. For mother's interoceptive awareness, the NPVI estimator found a significant effect whereas the linear estimator found no effect. We used simulations from empirical data to show that the NPVI estimator is more robust with respect to parametric assumption violations than linear estimators. Apart from the robustness, the NPVI estimator is also has the similar interpretability as linear coefficients. For example, The NPVI estimator for drive for thinness is 0.12(0.08, 0.15), which means if mother's drive for thinness is increased by 1, her child's BMI will increase by 0.12, with 95% confidence interval of (0.08, 0.15) (Table 4.2).

The data analysis and simulation results further illustrates the benefits of our approach in comparison to linear models and general machine learning models. The fact that our approach does not assume any form of relationship between all the variables make it robust with respect to model violations. On the other hand, both linear models and machine learning algorithms assume such a relationship and their corresponding variable importance measure reflects how well those models fit the data, not how changing the variable changes the outcome. Furthermore, our estimator is asymptotically linear (locally efficient) for which robust asymptotic inference is available. This makes our estimator interpretable to health care workers. For example, when the linear model assumptions hold true, our estimator has the same interpretability as the most widely accepted linear model coefficients when causality of the exposure on the outcome is established - both can be interpreted as: increasing the exposure by 1 will cause the outcome to increase by $\beta$).

# 4.6 Appendix

## Simulation Specification

**Algorithm 3** Linear Simulation Specification

1. Create $X_{Bin} = \mathbb{1}(X_0 == 0)$, and fit a logistic regression of $Y_0$ on $(X_{Bin}, W's)$, get the predicted probability of $Y_0 = 1$ as $g1w$;

2. For the vector of $W's$, generate a random variable $X$, where $X \sim Binomial(n, g1w)$ (n = number of observations in $W's$).
   **if** $X == 1$ **then**
   Fit a linear regression of $X_0$ on $W's$, get the predicted value $\hat{X}_0$, and the noise $\epsilon_X$ where $\epsilon_X \sim N(0, Var(\hat{X}_0 - X_0))$. Let $\hat{X} = \hat{X}_0 + \epsilon_X$.
   **else**
   $\hat{X} = 0$
   **end**

3. Fit a linear regression of $Y_0$ on $(\hat{X}, W's)$, get the predicted value $\hat{Y}_0$, and the noise $\epsilon_Y$ where $\epsilon_Y \sim N(0, Var(\hat{Y}_0 - Y_0))$. Let $\hat{Y} = \hat{Y}_0 + \epsilon_Y$.

4. Draw 1,000,000 samples of $W's$ with replacement randomly from the initial data set, use Step 1-3 to generate a simulated data set $df = (\hat{Y}, \hat{X}, W's)$. Calculate the asymptotic true value of $\Psi(P)$ as $\psi_0$ using proposition 1.

5. Draw $n$ samples of $W's$ with replacement randomly from the initial data set, generate simulated data set $df$, calculate the values of estimators $\Psi_{lm}(P_n)$, $\Psi_{SL}(P_n)$, $\Psi_{tmle.lm}(P_n)$, $\Psi_{tmle.npvi}(P_n)$ using $df$. Repeat for 1,000 times, and report the mean, MSE and coverage of 95% confidence intervals for the four estimators.

**Algorithm 4** Non-linear Simulation Specification

1. Create $X_{Bin} = \mathbb{1}(X_0 == 0)$, and fit a logistic regression of $Y_0$ on $(X_{Bin}, W's)$, get the predicted probability of $Y_0 = 1$ as $g1w$;

2. For the vector of $W's$, generate a random variable $X$, where $X \sim Binomial(n, g1w)$ (n = number of observations in $W's$).
   **if** $X == 1$ **then**
   Fit a linear regression of $X_0$ on $W's$, get the predicted value $\hat{X}_0$, and the noise $\epsilon_X$ where $\epsilon_X \sim N(0, Var(\hat{X}_0 - X_0))$. Let $\hat{X} = \hat{X}_0 + \epsilon_X$.
   **else**
   $\hat{X} = 0$
   **end**

3. Fit a Super Learner of $Y_0$ on $(\hat{X}, W's)$ (base learners include random forest and general additive models), get the predicted value $\hat{Y}_0$, and the noise $\epsilon_Y$ where $\epsilon_Y \sim N(0, Var(\hat{Y}_0 - Y_0))$. Let $\hat{Y} = \hat{Y}_0 + \epsilon_Y$.

4. Draw 1,000,000 samples of $W's$ with replacement randomly from the initial data set, use Step 1-3 to generate a simulated data set $df = (\hat{Y}, \hat{X}, W's)$. Calculate the asymptotic true value of $\Psi(P)$ as $\psi_0$ using proposition 1.

5. Draw $n$ samples of $W's$ with replacement randomly from the initial data set, generate simulated data set $df$, calculate the values of estimators $\Psi_{lm}(P_n)$, $\Psi_{SL}(P_n)$, $\Psi_{tmle.lm}(P_n)$, $\Psi_{tmle.npvi}(P_n)$ using $df$. Repeat for 1,000 times, and report the mean, MSE and coverage of 95% confidence intervals for the four estimators.

# Bibliography


[1] Paul W Holland. "Statistics and causal inference". In: *Journal of the American statistical Association* 81.396 (1986), pp. 945–960.

[2] Mark J van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

[3] Mark J Van Der Laan and Daniel Rubin. "Targeted maximum likelihood learning". In: *The international journal of biostatistics* 2.1 (2006).

[4] Megan S Schuler and Sherri Rose. "Targeted maximum likelihood estimation for causal inference in observational studies". In: *American journal of epidemiology* 185.1 (2017), pp. 65–73.

[5] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[6] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[7] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. "Super learner". In: *Statistical applications in genetics and molecular biology* 6.1 (2007).

[8] Anne Chao et al. "The applications of capture-recapture models to epidemiological data". In: *Statistics in medicine* 20.20 (2001), pp. 3123–3157.

[9] Stephen T Buckland et al. "Introduction to distance sampling: estimating abundance of biological populations". In: (2001).

[10] Mathew W Alldredge, Kenneth H Pollock, and Theodore R Simons. "Estimating detection probabilities from multiple-observer point counts". In: *The Auk* 123.4 (2006), pp. 1172–1182.

[11] Anne Chao. "An overview of closed capture-recapture models". In: *Journal of Agricultural, Biological, and Environmental Statistics* 6.2 (2001), pp. 158–175.

[12] Zachary T Kurtz. "Local log-linear models for capture-recapture". In: *arXiv preprint arXiv:1302.0890* (2013).

[13] Janet T Wittes. "Applications of a multinomial capture-recapture model to epidemiological data". In: *Journal of the American Statistical Association* 69.345 (1974), pp. 93–97.

[14] George Arthur Frederick Seber et al. "The estimation of animal abundance and related parameters". In: (1982).

[15] Hamparsum Bozdogan. "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions". In: *Psychometrika* 52.3 (1987), pp. 345–370.

[16] Gideon Schwarz et al. "Estimating the dimension of a model". In: *Annals of statistics* 6.2 (1978), pp. 461–464.

[17] David Draper. "Assessment and propagation of model uncertainty". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995), pp. 45–70.

[18] Emest B Hook and Ronald R Regal. "Validity of methods for model selection, weighting for model uncertainty, and small sample adjustment in capture-recapture estimation". In: *American journal of epidemiology* 145.12 (1997), pp. 1138–1144.

[19] Toon Braeye et al. "Capture-recapture estimators in epidemiology with applications to pertussis and pneumococcal invasive disease surveillance". In: *PloS one* 11.8 (2016), e0159832.

[20] Manjari Das and Edward H Kennedy. "Doubly robust capture-recapture methods for estimating population size". In: *arXiv preprint arXiv:2104.14091* (2021).

[21] Annegret Grimm, Bernd Gruber, and Klaus Henle. "Reliability of different mark-recapture methods for population size estimation tested against reference population sizes constructed from field data". In: *PLoS One* 9.6 (2014), e98840.

[22] Samuel G Rees et al. "Testing the effectiveness of capture mark recapture population estimation techniques using a computer simulation with known population size". In: *Ecological Modelling* 222.17 (2011), pp. 3291–3294.

[23] Robert C Spear et al. "Factors influencing the transmission of Schistosoma japonicum in the mountains of Sichuan Province of China". In: *The American journal of tropical medicine and hygiene* 70.1 (2004), pp. 48–56.

[24] René Gateaux. "Fonctions d'une infinité de variables indépendantes". In: *Bulletin de la Société Mathématique de France* 47 (1919), pp. 70–96.

[25] Richard D Gill, Jon A Wellner, and Jens Præstgaard. "Non-and semi-parametric maximum likelihood estimators and the von mises method (part 1)[with discussion and reply]". In: *Scandinavian Journal of Statistics* (1989), pp. 97–128.

[26] Mark J van der Laan, David Benkeser, and Weixin Cai. "Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso". In: *arXiv preprint arXiv:1908.05607* (2019).

[27] Richard M Cormack. "Log-linear models for capture-recapture". In: *Biometrics* (1989), pp. 395–413.

[28] Sophie Baillargeon, Louis-Paul Rivest, et al. "Rcapture: loglinear models for capture-recapture in R". In: *Journal of Statistical Software* 19.5 (2007), pp. 1–31.

[29] Zoe Emily Schnabel. "The estimation of the total fish population of a lake". In: *The American Mathematical Monthly* 45.6 (1938), pp. 348–352.

[30] *Schistosomiasis:Key Facts*. URL: `https://www.who.int/en/news-room/fact-sheets/detail/schistosomiasis`. (accessed: 04.07.2021).

[31] Song Liang et al. "Surveillance systems for neglected tropical diseases: global lessons from China's evolving schistosomiasis reporting systems, 1949–2014". In: *Emerging themes in epidemiology* 11.1 (2014), p. 19.

[32] Joseph L Doob. "The limiting distributions of certain statistics". In: *The Annals of Mathematical Statistics* 6.3 (1935), pp. 160–169.

[33] *Instituto Nacional de Salud Pública. Encuesta Nacional de Salud y Nutrición de Medio Camino 2016 (ENSANUT MC 2016)*. 2016.

[34] American Diabetes Association et al. "Standards of medical care in diabetes—2019 abridged for primary care providers". In: *Clinical diabetes: a publication of the American Diabetes Association* 37.1 (2019), p. 11.

[35] Joshua A Beckman and Mark A Creager. "Vascular complications of diabetes". In: *Circulation research* 118.11 (2016), pp. 1771–1785.

[36] Chin-Hsiao Tseng. "Mortality and causes of death in a national sample of diabetic patients in Taiwan". In: *Diabetes care* 27.7 (2004), pp. 1605–1609.

[37] Bianca Hemmingsen et al. "Targeting intensive glycaemic control versus targeting conventional glycaemic control for type 2 diabetes mellitus". In: *Cochrane Database of Systematic Reviews* 11 (2013).

[38] Antonio Méndez-Durán et al. "Current status of alternative therapies renal function at the Instituto Mexicano del Seguro Social". In: *Revista Médica del Instituto Mexicano del Seguro Social* 54.5 (2016), pp. 588–593.

[39] *Report to the Federal Executive and the Congress of the Union on the financial situation and risks of the Mexican Social Security Institute*. 2018.

[40] Ricardo Pérez-Cuevas et al. "Evaluating quality of care for patients with type 2 diabetes using electronic health record information in Mexico". In: *BMC medical informatics and decision making* 12.1 (2012), pp. 1–10.

[41] Svetlana V Doubova et al. "Loss of job-related right to healthcare is associated with reduced quality and clinical outcomes of diabetic patients in Mexico". In: *International Journal for Quality in Health Care* 30.4 (2018), pp. 283–290.

[42] Rafael Bustos-Saldaña et al. "Control de la glucemia en diabéticos tipo 2. Utilidad de mediciones en ayuno y posprandiales". In: *Revista Médica del Instituto Mexicano del Seguro Social* 43.5 (2005), pp. 393–399.

[43] Ana Marıa Salinas-Martınez et al. "Diabetes y consulta médica grupal en atención primaria:¿ Vale la pena el cambio?" In: *Revista médica de Chile* 137.10 (2009), pp. 1323–1332.

[44] Enrique Villarreal-Rıos et al. “Probability of control of the patient with diabetes exclusively treated with pharmacological therapy”. In: *Revista Médica del Instituto Mexicano del Seguro Social* 44.4 (2006), pp. 303–308.

[45] Edward H Wagner. “Chronic disease management: what will it take to improve care for chronic illness?” In: *Effective clinical practice* 1.1 (1998).

[46] Marco Antonio León-Mazón, Gerardo Jesús Araujo-Mendoza, and Zury Zaday Linos-Vázquez. “DiabetIMSS. Eficacia del programa de educación en diabetes en los parámetros clınicos y bioqumicos”. In: *Revista Médica del Instituto Mexicano del Seguro Social* 51.1 (2013), pp. 74–79.

[47] Ma Guadalupe Zuñiga-Ramirez et al. “Perfil de uso de los servicios del módulo DiabetIMSS por pacientes con diabetes mellitus 2”. In: *Rev Enferm Inst Mex Seguro Soc* 21.2 (2013), pp. 79–84.

[48] Marıa Eugenia Figueroa-Suárez et al. “Life style and metabolic control in DiabetIMSS program”. In: *Gaceta medica de México* 150.1 (2014), pp. 29–34.

[49] Judea Pearl et al. “Causal inference in statistics: An overview”. In: *Statistics surveys* 3 (2009), pp. 96–146.

[50] Scott L Zeger, Kung-Yee Liang, and Paul S Albert. “Models for longitudinal data: a generalized estimating equation approach”. In: *Biometrics* (1988), pp. 1049–1060.

[51] Stephen Milborrow et al. “earth: Multivariate adaptive regression splines”. In: *R package version* 5.2 (2017).

[52] Tianqi Chen et al. *xgboost: Extreme GradientBoosting*. 2018. URL: `https://CRAN.R-project.org/package=xgboost`.

[53] Ross Ihaka and Robert Gentleman. “R: a language for data analysis and graphics”. In: *Journal of computational and graphical statistics* 5.3 (1996), pp. 299–314.

[54] Eric Polley and Mark van der Laan. *SuperLearner: Super Learner Prediction*. Tech. rep. R package version 2.0-6. 2012. URL: `http://CRAN.R-project.org/package=SuperLearner`.

[55] Erin LeDell, Maya Petersen, and Mark van der Laan. “Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates”. In: *Electronic journal of statistics* 9.1 (2015), p. 1583.

[56] Susan Gruber and Mark J. van der Laan. “tmle: An R Package for Targeted Maximum Likelihood Estimation”. In: *Journal of Statistical Software* 51.13 (2012), pp. 1–35. URL: `http://www.jstatsoft.org/v51/i13/`.

[57] Leo Breiman et al. *Classification and regression trees*. CRC press, 1984.

[58] Alexander R Luedtke and Mark J Van Der Laan. “Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy”. In: *Annals of statistics* 44.2 (2016), p. 713.

[59] Lee Ling Lim et al. "Aspects of multicomponent integrated care promote sustained improvement in surrogate clinical outcomes: a systematic review and meta-analysis". In: *Diabetes Care* 41.6 (2018), pp. 1312–1320.

[60] Deise Regina Baptista et al. "The chronic care model for type 2 diabetes: a systematic review". In: *Diabetology & metabolic syndrome* 8.1 (2016), pp. 1–7.

[61] Edward H Wagner et al. "Improving chronic illness care: translating evidence into action". In: *Health affairs* 20.6 (2001), pp. 64–78.

[62] Brenda WC Bongaerts et al. "Effectiveness of chronic care models for the management of type 2 diabetes mellitus in Europe: a systematic review and meta-analysis". In: *BMJ open* 7.3 (2017), e013076.

[63] Pan American Health Organization. *Innovative Care for Chronic Conditions: organizing and delivering high quality Care for Chronic Noncommunicable Diseases in the Americas*. 2013.

[64] Julia Worswick et al. "Improving quality of care for persons with diabetes: an overview of systematic reviews-what does the evidence tell us?" In: *Systematic reviews* 2.1 (2013), pp. 1–14.

[65] American Diabetes Association et al. "12. Older adults: standards of medical care in diabetes—2019". In: *Diabetes Care* 42.Supplement 1 (2019), S139–S147.

[66] Ioannis Kavakiotis et al. "Machine learning and data mining methods in diabetes research". In: *Computational and structural biotechnology journal* 15 (2017), pp. 104–116.

[67] Pengfei Wei, Zhenzhou Lu, and Jingwen Song. "Variable importance analysis: A comprehensive review". In: *Reliability Engineering & System Safety* 142 (2015), pp. 399–432.

[68] Antoine Chambaz, Pierre Neuvial, and Mark J van der Laan. "Estimation of a non-parametric variable importance measure of a continuous exposure". In: *Electronic Journal of Statistics* 6 (2012), p. 1059.

[69] Mark J van der Laan and Susan Gruber. "Collaborative double robust targeted maximum likelihood estimation". In: *The international journal of biostatistics* 6.1 (2010).

[70] Mark J van der Laan and Susan Gruber. "Collaborative double robust targeted penalized maximum likelihood estimation". In: *UC Berkeley Division of Biostatistics Working Paper Series* (2009), p. 246.

[71] Susan Gruber and Mark J van der Laan. "An application of collaborative targeted maximum likelihood estimation in causal inference and genomics". In: *The International Journal of Biostatistics* 6.1 (2010).

[72] Alan E Hubbard, Chris J Kennedy, and Mark J van der Laan. "Data-adaptive target parameters". In: *Targeted Learning in Data Science*. Springer, 2018, pp. 125–142.

[73] Iván Dıaz et al. "Variable importance and prediction methods for longitudinal problems with missing variables". In: *PloS one* 10.3 (2015), e0120031.

[74] Alan Hubbard et al. "Targeted Learning for High-Dimensional Variable Importance". In: *University of California, Berkeley* (2016).

[75] Catherine Tuglus and Mark J van der Laan. "Targeted methods for biomarker discovery, the search for a standard". In: (2008).

[76] Hui Wang, Sherri Rose, and Mark J van der Laan. "Finding quantitative trait loci genes with collaborative targeted maximum likelihood learning". In: *Statistics & probability letters* 81.7 (2011), pp. 792–796.

[77] Hui Wang, Sherri Rose, and Mark J van der Laan. "Finding quantitative trait loci genes". In: *Targeted Learning*. Springer, 2011, pp. 383–394.

[78] Jeffrey A Kelly et al. "Community AIDS/HIV risk reduction: the effects of endorsements by popular people in three cities." In: *American Journal of Public Health* 82.11 (1992), pp. 1483–1489.

[79] David M Garner. *Eating Disorder Inventory-3 (EDI-3) Scale Descriptions*.

[80] Elissa S Epel et al. "The reward-based eating drive scale: a self-report index of reward-based eating". In: *PloS one* 9.6 (2014), e101350.

[81] Alison Gopnik, Laura Schulz, and Laura Elizabeth Schulz. *Causal learning: Psychology, philosophy, and computation*. Oxford University Press, 2007.

[82] EC Polley, AE Hubbard, et al. "Super learner." In: *Statistical applications in genetics and molecular biology* 6 (2007), Article25–Article25.

[83] Trevor Hastie. *gam: Generalized additive models*. version 1.15, 2018. URL: `https://CRAN.R-project.org/package=gam`.

[84] Charles Kooperberg. *polspline: Polynomial Spline Routines*. version 1.1.12, 2015. URL: `https://CRAN.R-project.org/package=polspline`.

[85] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In: *R News* 2.3 (2002), pp. 18–22. URL: `https://CRAN.R-project.org/doc/Rnews/`.