A New Method to Adjust for Ordinal Variables With a Pre-Diabetes Case Study

Chase Mathis[*],
Advisors: Robin J. Evans[†], Alexander Volfovsky[‡],
Committee: Peter Hoff[‡], and Surya T. Tokdar[‡]

[*]Trinity College of Arts & Sciences, Duke University, USA
[†]Department of Statistics, University of Oxford, UK
[‡]Department of Statistical Science, Duke University, USA

April 15, 2025

Abstract

Noisy measurements are ubiquitous in statistical analysis; recently methods to deal with these measurements have shown to work very well. One such field that has arisen is that of proximal causal inference. Proximal causal inference requires two proxies with specific assumptions about them. We consider the case of analyzing noisy measurements with only *one* proxy. Specifically, we will consider how to adjust for ordinal confounders in generalized linear models. We show that compared to common, naïve methods, our new method performs better across a variety of metrics in a simulation study. To illustrate the method, we re-analyze a previous study that used the NHANES public dataset. Software for the method and code to reproduce simulation results are publicly available.

1 Introduction

Noisy measurements have begun to become scrutinized in statistical methods such as multivariate regression. Modern datasets typically have many covariates, allowing the researcher to control for different types of variables and make exact statistical statements. Multivariable regression is a common statistical method in many applications. In the setting of causal inference in observational studies, controlling *exactly* for the variable is extremely important. Causal inference studies with observational data require the strong ignorability assumption: which cover uncofoundedness and positivity. Controlling for a noisy measurement is not sufficient to recover common causal inference estimands. Tchetgen Tchetgen and Miao have pioneered what is called proximal causal inference Tchetgen Tchetgen et al. (2020). Proximal causal inference places an emphasis on noisy measurements. Remarkably, with a noisy measurement associated with the treatment and a noisy measurement associated with the outcome, Tchetgen Tchetgen demonstrates the ability to use the two noisy measurements to recover a causal estimand. While this method works under broad distribution assumptions, we question the ability to extend this method to allow for only one proxy variable if we know the underlying distribution of the unobserved confounder.

In this project we develop a multivariable regression estimator for situations when one knows the true distribution but can only take noisy measurements from this distribution. The paper will focus on a specific example— where the noisy measurement is ordinal and the underlying distribution is normal, it is conceivable to extend this estimator to other distribution-measurement examples. Directed acyclic graphs (DAGs) will be used to visualize the causal assumptions and the proximal theorems that will be important for the estimator. To illustrate the novel estimator's improved performance, we do a simulation study for finite samples (n = 100, 1,000). Our evaluation metric for the simulation will be point estimate and false positive rates accuracy. In both metrics, the new estimator performs better than the naïve estimator most commonly used.

Finally, we apply our estimator to a cross sectional dentistry study, which uses the NHANES 2009-2010 survey to estimate how signs of periodontal disease is linked to pre-diabetes. We compare the researchers logistic regression conclusions with our new estimators. While the researchers concluded that there is a statistically significant impact of mild periodontal disease on prediabetes, our results do not show this effect.

Section 2 reviews current work on noisy measurement analysis and introduces preliminary definitions on causal inference estimands and DAGs. Section 3 describes our methodology, which we call the Bayesian Bootstrap. The methodology includes algorithms for logistic regression and linear regression, but can be extended to other generalized linear models by referencing Liu et al. (2024). Section 4 describes the details of our finite sample simulation study and subsequently illustrates the results. In section 5, we apply our method to our case study.

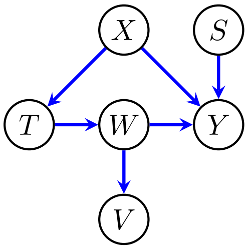
2 Background & Related Works

2.1 Causal Directed Acyclic Graphs

We begin with background on causal DAGs. A *graph* has a set of vertices $\mathcal{V}$ and a set of edges $\mathcal{E}$ of pairs of vertices. A *path* of length k is a sequence of $k+1$ distinct vertices. A path is called a cycle if it is of the form $v_0 \to \ldots \to v_k \to v_0$. An acyclic graph is a graph that has no cycles. Causal inference often works with these types of graphs and we will too. The following sets are important to know for causal DAG analysis: parents, ancestors, descendants, and causal nodes.

$$\begin{align}
\text{pa}_{\mathcal{G}}(v) &= \{w \in \mathcal{V} : w \to v \text{ in } \mathcal{G}\} \tag{1}\\
\text{an}_{\mathcal{G}}(v) &= \{w \in \mathcal{V} : w \to \ldots \to v \text{ in } \mathcal{G}\} \tag{2}\\
\text{de}_{\mathcal{G}}(v) &= \{w \in \mathcal{V} : v \to \ldots \to w \text{ in } \mathcal{G} \text{ or } w = v\} \tag{3}\\
\text{cn}_{\mathcal{G}}(v) &= \{w \in \mathcal{V} : \text{Treatment} \in \text{an}_{\mathcal{G}}(w) \text{ and Outcome} \in \text{de}_{\mathcal{G}}(w)\} \tag{4}
\end{align}$$

Notice in (4) that a vertex is its own descendant. There is also a notion of *strict* descendants defined as $\text{de}_{\mathcal{G}}(v) \setminus \{v\}$. In figure 1, we have a concrete visual example of these definitions and relationships within a DAG.



node	relationship	set
Y	$\text{pa}_{\mathcal{G}}$	$\{W, M, S\}$
V	$\text{an}_{\mathcal{G}}$	$\{W, T\}$
M	$\text{de}_{\mathcal{G}}$	$\{M, Y\}$
(T, Y)	$\text{cn}_{\mathcal{G}}$	$\{W, Y\}$

Figure 1: Example Graph

Statistically, the graphs verticies refer to random variables and the graphs edges refer to their associations. While all joint distributions can be written as a product of conditional distributions, the graph here allows the statistician to use conditional independence to write a more simplified joint distribution. In 1, the joint distribution can be written as

$$p(T, X, W, S, Y, V) = p(X)p(S)p(T \mid X)p(W \mid T)p(V \mid W)p(Y \mid S, X, W). \tag{5}$$

The ability to write out a joint in this form has many applications, including topics in predictive statistics named expert systems like junction trees and message passing (Pearl, 1988; Lauritzen and Spiegelhalter, 1988).

In addition to expert systems, decomposing the joint distribution into conditionals have applications in causal inference. For instance, in figure 1, we name three nodes T, X, Y specifically because these could be considered the treatment, a confounder, and an outcome in a causal inference analysis.

2.2 Causal Inference

In this section, we will introduce a slew of definitions and results that we will refer to throughout the development of our method. While we will use DAGs to visualize the conditional independencies and causal assumptions, we will use the potential outcomes framwork first formalized by Neyman in 1923 (Splawa-Neyman et al., 1990).

Definition 2.1. Average treatment effect Consider a treatment that takes on values $\{0, 1\}$ and an outcome Y. Consider n experimental units $\{Y_i\}_{1:n}$ (a person at a specific time, for instance). Denote $Y_i(j)$ to mean the outcome of the i^{th} experimental unit when given treatment j. We define the average treatment effect (also called average causal effect) to be

$$\tau \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - \frac{1}{n}\sum_{i=1}^{n} Y_i(0). \tag{6}$$

While this definition is defined on a binary treatment, the definition can be extended to discrete or continuous treatments by formulating a model and estimating the expected change in outcome. The fundamental problem in causal inference is that we cannot observe both treatments: $Y_i(1)$ and $Y_i(0)$. While there is some research focusing on causal inference without counterfactual Dawid (2000), most theory surrounding causal inference assumes this counterfactual structure.

A classical method of estimating a treatment effect is through a randomized experiment. While in this work, we will not discuss the importance of randomization (Fisher Sharp Null, Neyman Null, Stratification, etc.) we pause and question why randomization is needed. One reason we cannot just *observe* people who have taken the treatment and people who have not and compute a simple estimator is because of confounders.

Definition 2.2. Confounder A confounder is a random variable X such that the association between two random variables (Y, Z) change if we know the value of X (Vander Weele and Shpitser, 2011). This is formally known as the Yule-Simpson paradox (Simpson, 1951; Yule, 1902).

Graphically, a confounder is a variable that causes both a treatment and an outcome. For instance in figure 1, X is a confounder with respect to (T, Y).

In this vein of relating mathematics to graphical models, a method of translating a graph to a set of mathematical distribution assumptions is through a structural equation model (SEM) (Wright, 1921).

Definition 2.3. Structural Equation Model (SEM) For a graph $\mathcal{G}$ and probability distribution p, we say that $(\mathcal{G}, p)$ is a structural equation model if for each vertex $j \in \mathcal{V}$

$$X_j = \left(\sum_{i \in \mathrm{pa}_{\mathcal{G}}(j)} b_{ij} X_i \right) + d_{jj} \tag{7}$$

Note that in (5), we were not assuming that the probability measure followed a structural equation model. We only assumed the graph was indicating to us notions of conditional independence. Therefore, each variable is a linear combination of its parents, plus some independent residual variance. A multivariate normal distribution can always be written as (7) (Lauritzen, 1996). Note that we write b_{ij} (not β_{ij}) for the edge weights between vertex i to vertex j and d_{ii} for the internal variance of node i.

If there exists a probability distribution p such that with our example graph in figure 1 is a structural equation model, we would write, for example:

$$Y = b_{MY} \times M + b_{WY} \times W + b_{SY} \times S + d_{YY}.$$

2.3 Proximal Causal Inference and Noisy Measurements

Confounding is a core reason association does not imply causality. If we know which variables are confounding the $(T \to Y)$ causal effect, there are many methods to control or remove this nuisance parameter (Robins, 1986). Unfortunately, there exist many scenarios where we cannot measure all of the confounders. Measuring more covariates can reduce the risk of unobserved confounders. But it can be expensive or even impossible to measure all confounder, we may want to control for in a statistical analysis.

In these situations, researchers must deal with what we call an unobserved confounder. Problems of unobserved confounding are very important to statistics research, resulting in the following method's authors receiving the Nobel prize in 2021—the instrumental variable method (Angrist et al., 1996). The method relies on the existence of an instrumental variable (IV). An IV is a covariate that causes the treatment, and satisfies for T, U, Y the treatment, confounder, and outcome $I \perp\!\!\!\perp Y$ and $I \perp\!\!\!\perp Y \mid (U, T)$.

The graph below illustrates the set of basic assumptions for an instrumental variable model. We have our instrumental variable I causes T. T causes Y and U is an unobserved confounder—U is a common cause of T and Y.

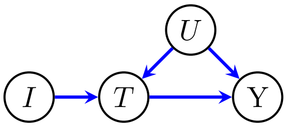


Figure 2: Instrumental Variables

For notational purpose, we define lm to the Ordinary Least Squares estimator in linear regression. The algorithm is a simple two stage least squares regression. First we calculate the fitted values of the regression of T by the IV I. Using this fitted values, we run a second regression of Y by the fitted values $\widehat{T}$. The intuition behind this method is that we are isolating the information in T that cannot be explained by U. A summary of the algorithm can be found in algorithm 1.

Algorithm 1 Two Stage Least Squares Instrumental Variable Algorithm under Linearity Assumption

```
Input: Feature matrix $\mathbf{X} = \begin{pmatrix} I & T & U & Y \end{pmatrix}$
Output: ATE $\tau$
$\widehat{T} \leftarrow lm(T \sim I)$
$\tau \leftarrow lm(Y \sim \hat{T})$ the coefficient estimated for T.
return $\tau$
```

This algorithm is commonly referred to as a "two stage" least squared (2SLS) because of the two stages of ordinary least squares (OLS) estimation.

Theorem 2.4. ***Instrumental Variable Identification Algorithm Works*** *Algorithm 1 gives an unbiased estimate $\hat{\tau}$ for our statistics τ.*

2.4 Proximal Causal Inference for Continuous Y and Continuous W

We introduce instrumental variable identification because its close relationship to a novel set of methods under the name proximal causal inference. Both instrumental variable method and proximal causal inference methods are able to estimate the ATE τ with the presence of unobserved confounders. Figure 3 illustrates the set of assumptions for proximal causal inference through a causal DAG.

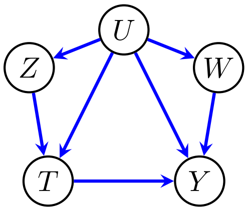


Figure 3: Allowed Causal DAG for Proximal Causal Learning under continuous Y and continuos W.

Again, (T, Y) is our treatment outcome pair, U is an unobserved confounder and now (Z, W) are the treatment proxy and outcome proxy respectively. With this information, we can recover the ATE τ with the following algorithm, Tchetgen Tchetgen calls Proximal 2SLS (P2SLS). The P2SLS algorithm for linear outcomes and proxy W is similar to that explained in algorithm 1. First, we fit a linear model predicting W from treatment and treatment proxy T, Z. Then using the fitted values $\widehat{W}$, we fit a regression Y by $T + \widehat{W}$ and recover τ. A summary of the algorithm is found in algorithm 2.

Algorithm 2 Proximal Two Stage Least Squares Algorithm

```
Input: Feature matrix $\mathbf{X} = \begin{pmatrix} Z & W & T & U & Y \end{pmatrix}$
Output: ATE $\tau$
$\hat{W} \leftarrow lm(W \sim T + Z)$
$\tau \leftarrow lm(Y \sim T + \hat{W})$ the coefficient for $T$.
return $\tau$
```

Theorem 2.5. ***The P2SLS Algorithm Works***

Proof For this project, I will only share the proof of the linear case (where conditional expectation of Y, W are linear). This proof is the most illuminating and the easiest to follow. The proof is taken from recent work on regression based proximal causal inference by Liu et al. (2024). For the most general assumptions, see (Tchetgen et al., 2020).

Assumptions:

$$\begin{aligned}
&E[Y \mid T, Z, U] = \beta_0 + \beta_T \times T + \beta_U \times U \\
&E[W \mid T, Z, U] = \alpha_0 + \alpha_U \times U \\
&|E[U \mid T, Z]| < \infty
\end{aligned}$$

Ultimately, we want to find $E(Y \mid T, Z)$ unconditional on U. To do this, first find

$$E(W \mid T, Z) = E\left(E(W \mid T, Z, U) \mid T, Z\right) = \alpha_0 + \alpha_U E(U \mid T, Z) \tag{8}$$

Now, we calculate

$$E(Y \mid T, Z) = E\left(E(Y \mid T, Z, U) \mid T, Z\right) = \beta_0 + \beta_T \times T + \beta_U E(U \mid T, Z) \tag{9}$$

Re-arranging equation (8), we get

$$E(U \mid T, Z) = \frac{E(W \mid T, Z) - \alpha_0}{\alpha_0} \tag{10}$$

allowing us to plug in this value to equation 9

$$E(Y \mid T, Z) \propto \beta_T \times T + \beta_U \times E(W \mid T, Z)$$

2.5 Proximal Causal Inference for Binary Y

Many times in causal inference and in the case study we share at the end, the outcome is binary. For a binary outcome, the causal assumptions differ from that of the linear model. Namely instead of $W \to Y$ in the linear case, now $Y \to W$. We think the differing assumptions is an interesting starting point of future research in this topic.

Specifically, the assumptions are

$$\begin{aligned}
&E[\text{logit}(\text{P}(Y)) \mid T, Z, U] = \beta_0 + \beta_T \times T + \beta_U \times U \\
&E[W \mid T, Z, U, Y = 1] = \alpha_0 + \alpha_U \times U + \alpha_Y \times Y \\
&|E[U \mid T, Z, Y]| < \infty
\end{aligned}$$

We can summarize these assumptions in figure 4.

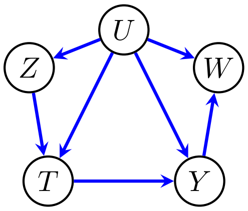


Figure 4: Allowed Causal DAG for Proximal Causal Learning and Continuous Outcomes. Note that $Y \to W$. Previously, $W \to Y$.

For notational purposes, we define *logitm* to be the logistic regression model.

Algorithm 3 Proximal Two Stage Least Squares Algorithm for Binary Y Continuous W

```
Input: Feature matrix $\mathbf{X} = \begin{pmatrix} Z & W & T & U & Y \end{pmatrix}$
Output: ATE $\tau$
$\hat{\beta} \leftarrow lm(W \sim T + Z + Y)$
$\widehat{W} \leftarrow (1, T, Z, 1)\hat{\beta}$
$\tau \leftarrow logitm(Y \sim T + \widehat{W})$ the coefficient for $T$.
return $\tau$
```

The proof for this model assumption is much longer and can be found in Liu et al. (2024). One important point of contrast with this algorithm compared to the linear algorithm is that $\widehat{W}$ is not just the fitted values, but instead we condition on $Y = 1$.

2.6 Probit Regression for Ordinal Data

The last preliminary we will need for our method is understanding probit regression for ordinal data. First, we note that *probit regression* is based upon using the normal cumulative distribution function, Φ, to transform numbers in $\mathbb{R}$ to those in $[0, 1]$. Second, we say that data is *ordinal* if it is finite and the ordering encodes important information. An example of ordinal data could be the following question on a health survey

How often do you exercise: a: Never, b: Rarely, c: Sometimes, d:Often, e: Very Often

whereas an example of a finite sized data that is not ordinal could be the answer to this survey question

Which ride sharing company are you most familiar with: a.) Uber b.) Lyft c.) Waymo

Bayesian method:

The latent-variable probit ordinal regression model is specified in (Hoff, 2009) as

$$\varepsilon_1, \ldots, \varepsilon_n \sim N(0, 1)$$

$$Z_i = \beta^T X_i + \varepsilon_i$$

$$O_i = g(Z_i)$$

where $\boldsymbol{\beta}, g$ are random parameters and g is monotonically increasing. In our method, we will want to be able to simulate Z_i conditional on observing the ordinal data O_i. To do this, we can calculate the full conditional

$$p(Z_i \mid \boldsymbol{\beta, O, g}) \propto \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z_i - \mathbf{X_i}\beta)\right) \times \delta_{a,b}(z_i) \tag{11}$$

The full conditional in equation (11) allows us to sample draws from the posterior distribution of Z_i if we have an estimated $\boldsymbol{\beta}, \boldsymbol{g}$ as $\boldsymbol{O}$ is observed. The name of this distribution is a constrained normal. We refer as the distribution with density proportional to 11 as $\mathrm{CN}(\boldsymbol{\beta}^{\mathrm{T}}\mathrm{X_i}, \mathrm{O_i}, \mathrm{g})$ as $\delta_{a,b}$ can be calculated from g and O. A method of simulating from this simulation can be found in (Hoff, 2009).

There are a few methods to estimate $\boldsymbol{\beta}, \boldsymbol{g}$. The first is a Bayesian method, where one can specify their priors and use a Gibbs-sampler to estimate the β parameter.

The priors can specified as

$$\boldsymbol{\beta} \sim N(\mu_{0,\beta}, \sigma^2_{0,\beta})$$

$$\alpha \sim N(\mu_{0,\alpha}, \sigma^2_{0,\alpha})$$

Then,

$$Z_i = \boldsymbol{\beta}^T X_i + \alpha + \varepsilon_i$$

$$c_i \sim N(k_i, 1)$$

and the likelihood can be taken from the Stan (Gabry et al., 2024) package called the ordered_logistic likelihood. Mathematically, the orderered logistic likelihood is the function

$$\text{OrderedLogistic}(k \mid \eta, c) = \begin{cases} 1 - \text{logit}^{-1}(\eta - c_1) & \text{if } k = 1, \\ \text{logit}^{-1}(\eta - c_{k-1}) - \text{logit}^{-1}(\eta - c_k) & \text{if } 1 < k < K, \text{and} \\ \text{logit}^{-1}(\eta - c_{K-1}) - 0 & \text{if } k = K. \end{cases}$$

The advantages of using this method are the ability to set priors and subsequently the ability to estimate uncertainty under these priors. Moreover, if data is sparse this method can be superior. The disadvantages are having to use Markov Chain Monte Carlo (MCMC). While for any single use, the MCMC is valid for researchers, in a simulation study the MCMC takes too long. Therefore, in the simulation study we will describe in 4, we use the maximum likelihood based Cumulative Linked Model that assumes

$$\mathrm{logit(P(O_i \leq j) = -\boldsymbol{\beta}^T X - \beta_0}$$

where the negative signs are for convenience as positive coefficients are related to increases in category. We use the ordinal package to fit this model(Christensen, 2022).

Assume we have fitted one of these two models above and have estimates for $\hat{\boldsymbol{\beta}}, \hat{\alpha}$, we can simulate from the posterior predictive distribution as follows.

Algorithm 4 Simulating Underlying Normal Process from Posterior Predictive

```
Input: Explanatory Variables $\mathbf{X} = \big(X_1, \ldots, X_k\big)$, Ordinal data $\boldsymbol{Y}$, Estimated parameters $\hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\boldsymbol{g}}$
Output: $\boldsymbol{Z}$ the latent variable that generates $\boldsymbol{Y}$
for $j$ in 1:n do
    $Z_j \sim CN(\boldsymbol{\beta}^T X_i, O_i, g)$
end for
return $\{Z_i\}_{i=1:n}$
```

3 Methodology

In the Proximal Causal Inference framework, two proxies are required for the absence of one confounder. Our methodology will extend this framework to allow for only **one** proxy variable through a Bayesian sampling procedure. We call this method the Bayesian Bootstrap for Proximal Causal Inference.

3.1 Bayesian Bootstrapping for One (1) noisy measurement

Consider the causal graph below and note we no longer observe the confounder. We often call this variable U, standing for "unobserved". This variable is often called a "hidden state" in Hidden Markov Models or simply an unobserved confounder in causal settings.

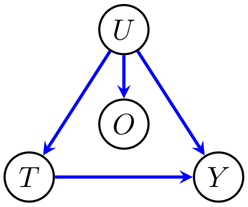


Figure 5: The starting causal graph.

Our causal estimated going forward will be the ATE (see equation 6). If we want to estimate the ATE, we must control for U Pearl (1993). U is an unobserved random variable, so we cannot control for it. One may and often does control for O (Arora et al., 2014). Controlling for O will not produce a consistent or unbiased estimator for the ATE, which begs the question: Is it possible to recover a better estimate of U we can control for?

Main Result1: Bootstrapping Proxies for Continuous Y If we assume (U, Y) are jointly normal, then we can use probit regression to simulate values of U conditional on either T, or Y, or both using its posterior distribution where the posterior is a constrained normal distribution [see Hoff (2009) for more].

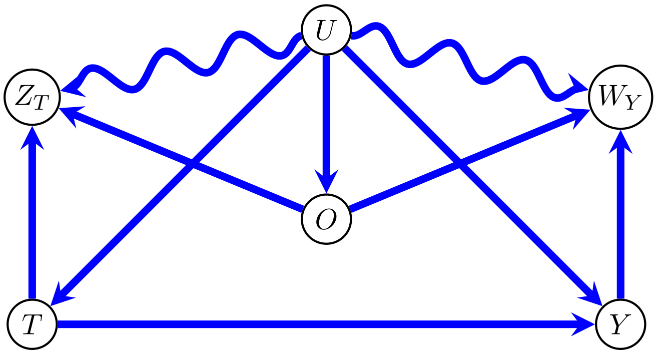


Figure 6: Bayesian Bootstrap Causal DAG. Here, Z_T is the simulations from T and W_Y is the simulations from Y.

To transform the graph seen in 5 to figure 6 we fit two separate Bayesian models to get two different latent variables.

$$Z_T \leftarrow f(T, O)$$

and

$$W_Y \leftarrow f(Y, O)$$

where f is the function defined by algorithm 4.

Naturally T generates Z_T and because we are using O as our ordinal data O helps in generating Z_T. The remarkable thing is empirically, $Z_T \not\perp\!\!\!\perp U \mid (T, O)$. We are getting *new* information about U that is not in O. The same property is observed in W_Y. This property may seem like "magic" as we are gaining new information outside of our original dataset. I conjecture because we are assuming the normal process in our U and in the Bayesian model, we are able to get this extra partial correlation. A detail to notice in the diagram is $W_Y \in \text{de}_{\mathcal{G}}(Y)$. In the continuous case, our assumption described in figure 3 is violated the resulting estimator will be biased. Our method finds that empirically the bias is small, so for the continuous case we continue with the biased estimator. This is a major point of future research that can allow the proximal two stage least squares to be more flexible. However, if Y is binary, algorithm 3 should be unbiased. This leads to our second main result.

Main Result 2: Bootstrapping Proxies for Binary Y Motivated by our case study Arora et al. (2014), we assume a slightly more complicated causal set up. In many research applications a binary variable is generated by a continuous variable. For instance, our case study example has a pre-diabetes outcome where the definition of pre-diabetes is an indicator function of blood sugar levels. This motivates the use of an equivalent definition of logistic regression. This equivalent definition is the latent-variable definition where

$$Y_i^* = \beta_T \times T_i + \varepsilon_i$$

where $\varepsilon_i \overset{i.i.d.}{\sim}$ Logistic$(0, 1)$ and the binary variable is

$$Y = \mathbb{I}_{Y*>0}.$$

Therefore, if the researcher does not have access to the latent variable generating the binary variables this method should **not** be used. In addition, if T is binary or ordinal, but the researcher has access to the full data, we can employ a similar technique. The causal assumptions for these situations are summarized in figures 7 and 8

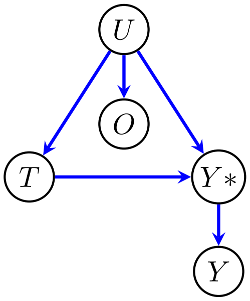


Figure 7: Assumed causal DAG for Y outcome.

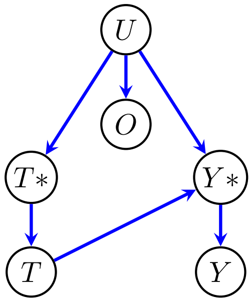


Figure 8: Assumed causal DAG for discrete (T, Y)

Similar to the continuous case, we generate

$$Z_T \leftarrow f(T, O)$$

and

$$W_Y \leftarrow f(Y*, O).$$

Therefore, everything between the binary case and the continuous case is the same, except for the fact that we use the latent variable to generate the bootstrapped proxies. One can deduce there is little predictive power in the binary data, because there is much less information in zero-one data as there is in the latent variable that generates it. Moreover, in almost all situations the researcher has access to this latent variable as they do in our case study. Therefore, we believe these set of assumptions are reasonable. Focusing for simplicity on the assumptions laid out in 7, we get the causal structure in figure 9 after generating Z_T, W_Y. Once we have this structure, we use algorithm 3 to get the coefficient of $T \rightarrow Y$ adjusting for coefficient.

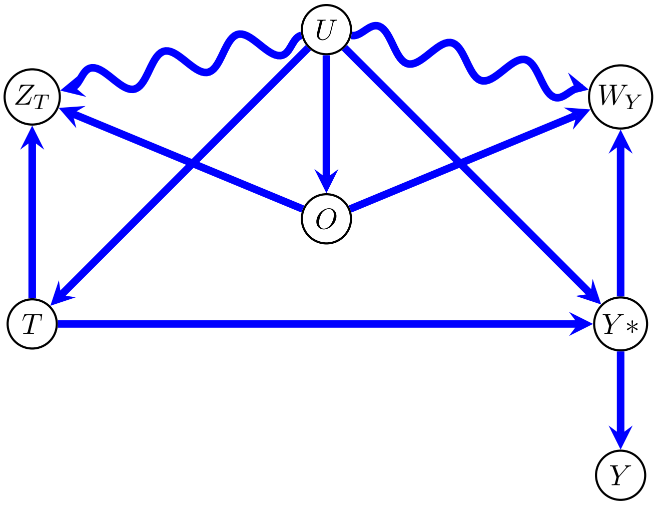


Figure 9: Bayesian Bootstrap Causal DAG. Here, Z_T is the simulations from T and W_Y is the simulations from Y.

4 Simulation Study

Our results, which take form of a finite sample simulation study, can be summarized in two categories. First, we calculate the accuracy of the actual point estimate generated by our Bayesian bootstrap method. Secondly, we perform a sensitivity analysis of our method by considering the false positivity rate across the different methods. In both point estimates and type 1 error rate calculations, the Bayesian bootstrap method performs superior to the naive calculation of controlling for the ordinal variable itself.

4.1 Simulation details

First, we will discuss the details of the simulation set up. Simulations for causal inference estimators are very common, but setting up the simulation must be a careful design to allow the method to be exported to real world settings. For instance, varsortability (Reisach et al., 2021) is a warning when conducting simulations. The warning is directed to researchers attempting to recover the DAG—as nodes with higher variances are likely to have many ancestor nodes (recall 2). Our method does not attempt to learn DAG structure, although the warning indicates that simulating DAGs can have non-uniform correlation structures. We take this warning and proceed very carefully as described in the following paragraph.

Consider the graph in figure 6. For each blue arrow in the figure, we sample an edge coefficient uniformly at random $\in [-5, 5] \subset \mathbb{R}$. In addition for each node in the graph, we will sample dispersion or residual variance parameters uniformly at random $\in [1, 10] \subset \mathbb{R}$. We make sure to sample smaller magnitude edge-coefficients to reduce this issue of non uniform correlation structure as described in the varsortability paper (Reisach et al., 2021). In a similar vein, we calculate and keep track of a signal-to-noise statistic. The signal-to-noise statistic is calculated as the partial correlation of T, Y conditional on U which we write as $\rho_{TY|U}$. We do this to narrow down which situations the method works the best. In our results, we define the categories

$$h(\rho_{TY|U}) = \begin{cases} \text{Low} & \left|\rho_{TY|U}\right| < 0.25 \\ \text{Medium} & 0.25 < \left|\rho_{TY|U}\right| < 0.75 \\ \text{High} & 0.75 < \left|\rho_{TY|U}\right| \end{cases} \tag{12}$$

A summary of the simulation algorithm is shown below for continuous Y.

Algorithm 5 Simulating Details for Simulation Study of Bayesian Bootstrap for Continuous (T, Y)

```
Input: Number of data points $n$
Output Simulated Dataset: $\boldsymbol{X}$
$dd_U, dd_Y, dd_T \overset{i.i.d.}{\sim} U(1, 10)$
$b_{UT}, b_{UY}, b_{TY}, b_{UO} \overset{i.i.d.}{\sim} U(-5, 5)$
for i in 1:n do
    $U_i \sim N(0, dd_u)$ i
    $T_i \sim N(b_{UT} \times U_i, dd_T)$
    $Y_i \sim N(b_{UY} \times U_i + b_{TY} \times T_i, d_{YY})$
    $\eta_i = b_{UO} \times U_i$
    $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are equidistant in $(\min(\eta_i), \max(\eta_i))$
    $p_{ik} = \text{logit}(\alpha_\mathrm{k} + \eta_\mathrm{i})$
    $O_i = 1 + \sum_{k=1}^{4} \mathbb{I}_{W > p_{ik}}$ where $W \sim U(0, 1)$
end for
return $X = \begin{bmatrix} U_i & T_i & Y_i & O_i \end{bmatrix}$
```

If T is binary and Y is binary we have

Algorithm 6 Simulating Details for Simulation Study of Bayesian Bootstrap for Binary T, Y

```
Input: Number of data points $n$
Output Simulated Dataset: $\boldsymbol{X}$
$dd_U, dd_Y, dd_T \overset{i.i.d.}{\sim} U(1, 10)$
$b_{UT}, b_{UY}, b_{TY}, b_{UO} \overset{i.i.d.}{\sim} U(-5, 5)$
$\varepsilon_{i,T,Y} \overset{i.i.d.}{\sim} \text{Logistic}(0, 1)$
for i in 1:n do
    $U_i \sim N(0, dd_u)$ i
    $T*_i \leftarrow b_{UT} \times U_i + \varepsilon_{i,T}$
    $T_i \leftarrow \mathbb{I}_{T*_i} > 0$
    $Y*_i \leftarrow b_{UY} \times U_i + b_{TY} \times T_i + \varepsilon_{i,Y}$
    $Y \leftarrow \mathbb{I}_{Y*_i} > 0$
    $\eta_i = b_{UO} \times U_i$
    $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are equidistant in $(\min(\eta_i), \max(\eta_i))$
    $p_{ik} = \text{logit}(\alpha_\mathrm{k} + \eta_\mathrm{i})$
    $O_i = 1 + \sum_{k=1}^{4} \mathbb{I}_{W > p_{ik}}$ where $W \sim U(0, 1)$
end for
return $X = \begin{bmatrix} U_i & T*_i & T_i & Y*_i & Y_i & O_i \end{bmatrix}$
```

For our simulation study, we use three different types of estimators. The **first** is our proposed Bayesian Bootstrap method where we first simulate according to equation (4) and then use P2SLS to estimate $\hat{\beta}_{TY}$ according to algorithm 2 (or algorithm 3 for binary Y). The second estimator we keep track of the *oracle estimator*, which is the regression estimate $\hat{\beta}$ for $y = Y, X = [T, U]$. It is called oracle because it has access to the unobserved data.

Finally, we keep track of what we call the "naïve" estimator. While we call this estimator the naïve estimator, it is a complicated method not often taught to statisticians in undergraduate [1].The naïve method we use is the polynomial contrast (Montgomery, 2013).

4.2 Simulation Results for Continuous (T, Y)

Using algorithm 5, we simulate a total of 20,000 datasets where we assume (T, Y) are continuous. 10,000 of these datasets are smaller with $n = 100$ and 10,000 of these datasets are medium sized with $n = 1,000$.

The error metric we choose is the squared error given by

$$(b_{TY} - \hat{\beta}_{TY})^2$$

where b_{TY} is the true edge coefficient and the estimate comes from which estimator we use.

As the section header announces, we display and discuss the results for when T, Y are continuous. The results are very similar to when T, Y are binary and do not need further discussion. Therefore, we place these results in the appendix section 6.

4.2.1 Point Estimate Accuracy

Number of Samples	Mean Error	Median Error	Simulation Type
100	0.2320	0.0343	Bayesian Bootstrap
1000	0.2098	0.0168	Bayesian Bootstrap
100	0.3530	0.0799	Naive
1000	0.3149	0.0622	Naive
100	0.0148	0.0043	Oracle
1000	0.0026	0.0004	Oracle

Table 1: Improved point estimate accuracy when using a Bayesian Bootstrap

Table 6 illustrates that the average squared error and the median squared error is smaller when using our proposed Bayesian Bootstrap method as compared to the naïve method. The improvement is true across small and medium datasets, though as n grows the improvement becomes smaller. While this table shows that overall, our proposed method is better than the naïve method commonly used, we wish to analyze under what situations does this method improve the most. Therefore, in figure 10, we display the improvement stratified by our partial correlation categories defined in equation (12).

[1] We tested a myriad of different methods to control for O "naïvely" and all had similar results

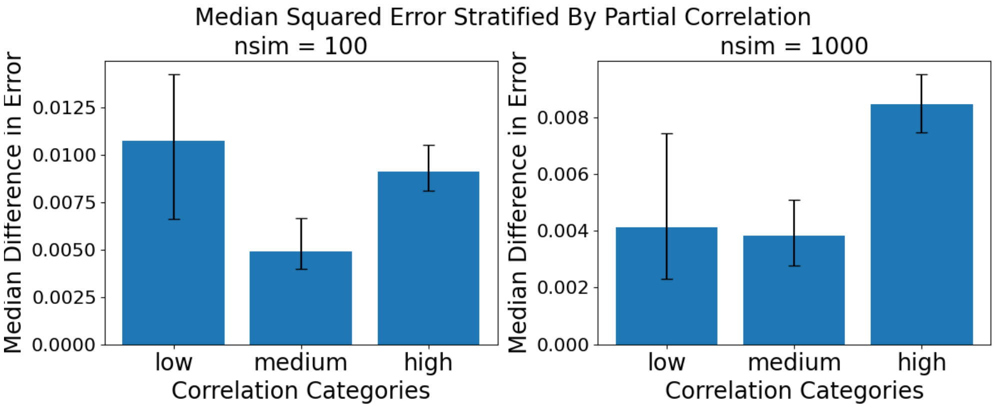


Figure 10: Median Squared Error improvement with Bayesian Bootstrap compared to Naive Estimator. There does not exist a clear trend across signal to noise. In fact, several of the confidence intervals overlap with each other.

4.2.2 Type 1 Error Rate Comparison

Type 1 error rate is incredibly important in applied statistics. Many scientific findings main results are related to reporting significant effects in regression tables and simple hypothesis tests. In our simulation result, we find that the type 1 error decreases when using the Bayesian bootstrap as compared to controlling for the ordinal variable naively. For this result, the simulation details remain the same, but we set b_{TY} to zero for all simulations.

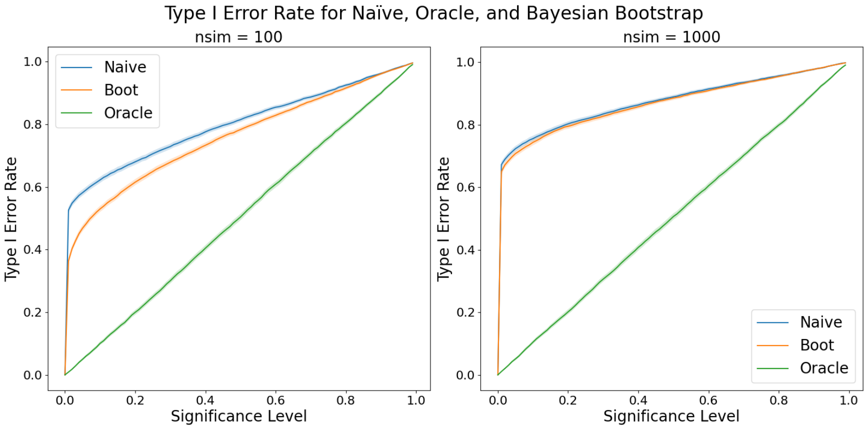


Figure 11: Type 1 error rate comparison with Bootstrap technique performing better than the naive adjustment.

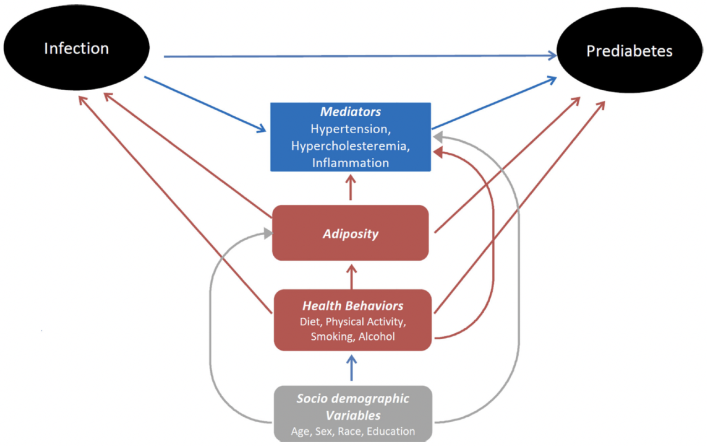
For each dataset our estimator sees, we calculate a point estimate and a p-value using the lm function in R (R Core Team, 2023). We keep track of these reported p-values and see the percent of the p-values that are less than a specified α level on the x-axis. Because the oracle is specified completely correctly, p-values are distributed $U(0, 1)$, meaning α percent of them are less than α. While the bootstrap method is not perfect, the method improves its type 1 error rate compared to the naïve model.

5 Case Study: Periodontal Disease and Prediabetes

Researchers focused on dental health were interested in the analyzing the relationship between periodontal disease (a gum disease) and prediabetes (Arora et al., 2014). The study analyzed cross sectional data from the National Health and Nutrition Examination Survey (NHANES) 2009-2010 publicly available dataset to analyze this relationship. The researchers were very careful in their statistical methods, making their assumptions clear, using the recommended survey weights from the NHANES dataset, excluding blatant outliers from the data, and isolating the relationship between periodontal disease and prediabetes through a logistic regression. The conclusions of their study can be summarized by this quotation:

> “Periodontal infection was positively associated with prevalent impaired glucose tolerance in a cross-sectional study among a nationally representative sample.”

The results come from a logistic regression, where they control for many confounders they assumed. They include a DAG in their analysis to visualize their assumptions (see fig 5).



Unfortunately, the researchers did not make code available for this study. While replicating the study, the summary statistics all matched the researchers reported summary statistics, but the regression point estimates did not exactly correspond to reported estimates. However, confidence regions matched mostly and significant results remained significant under replication. We present two comparisons using the replicated results as comparison. We also attach the reported estimates in the appendix section 6. The researchers tested multiple types of models, but ultimately ended up using a model that controlled for socio demographic variables, health behaviors, and adiposity (Body mass index). The estimates can be found in table 2.

	Periodontal Disease Severity	
Model	Moderate Periodontis	Severe Periodontis
Oral Glucose	1.23 (0.59–2.59)	**2.35 (1.26–4.37)**
Impaired Fasting Glucose	1.17 (0.76–1.81)	1.31 (0.71–2.43)

Table 2: Odds ratios (95 % CI) for the association between periodontal disease severity and glycemic outcomes using the naïve method to control for ordinal variables. Controlling for socio demographic variables, health behaviors, and adiposity (Body mass index).

Education is an ordinal variable, which implies the ability to use the Bayesian Bootstrap method. Therefore, when one uses the Bayesian Bootstrap method, the estimates become

	Periodontal Disease Severity	
Outcome	Moderate	Severe
Oral Glucose	1.19 (0.56–2.54)	**2.02 (1.03–3.96)**
Impaired Fasting Glucose	1.17 (0.75–1.81)	1.21 (0.63–2.32)

Table 3: Odds ratios (95% CI) for the association between periodontal disease severity and glycemic outcomes using Bayesian Bootstrap. Controlling for socio demographic variables, health behaviors, and adiposity (Body mass index).

Here, the Bayesian bootstrap method shrinks the naïve estimators so that the link between severe periodontal disease and oral glucose is much less strong and less statistically significant. We note that the p-value reported for the naïve estimator is *0.01*, whereas the p-value reported for the novel method is *0.04*.

Our second comparison is for a supplementary model the authors decided to test. In this model, the researchers only control for socio demographic variables. The estimates can be found in table 4.

	Original Analysis	
Outcome	**Moderate**	**Severe**
Oral Glucose Tolerance	1.20 (0.486–2.95)	**2.08 (1.01–4.29)**
Impaired Fasting Glucose	1.14 (0.680 – 1.90)	1.28 (0.660–2.47)

Table 4: Odds ratios (95 % CI) for the association between periodontal disease severity and glycemic outcomes using the naïve method to control for ordinal variables. Only controlling for socio-demographic variables.

Similar to our prior comparison, the estimates in table 4 uses a naïve method to control for education, whereas estimates in table 5 use the Bayesian bootstrap method.

	Periodontal Disease Severity	
Outcome	Moderate	Severe
Oral Glucose	1.14 (0.52–2.50)	**1.72 (0.83–3.56)**
Impaired Fasting Glucose	1.11 (0.71–1.74)	1.11 (0.62–1.98)

Table 5: Odds ratios (95% CI) for the association between periodontal disease severity and glycemic outcomes using Bayesian Bootstrap. Controlling only for socio demographic variables.

In this comparison, our method fails to find a link between periodontal infection and prediabetes, where the researchers did find an infection. With our simulation study results in section 4, we believe the authors could have made a type 1 error.

6 Discussion

In this paper, we developed a regression method when there exists ordinal confounders. The method improves upon many metrics when compared to the naïve method most regressions use. In particular, the Bayesian Bootstrap method has smaller point estimate accuracy and reduces the type 1 error rate. The results of this have been validated by a large simulation study on multiple data generating processes. While the method assumes only ordinal data and normal underlying process, it is not hard to imagine future additions where those details can be changed. We hope our results add not one method in a specific context, but rather motivate new ideas on how to use all the information in the data.

One limitation of this method is that it it relies on many distribution assumptions. While this is most definitely an area for improvement, we believe the methods advantage relate to the disadvantages of using classical proximal causal inference. Proximal causal inference requires the researcher to find *two* proxies and each of them are not confounders themselves. While the former requirement may be easy, the later is not. For instance, we may want to use gender as a proxy for education, but that is assuming that gender itself is not a confounder. Here, we are only using causal objects we already specified and do not make additional assumptions.

We summarize future research methods we think are interesting to follow

- In the simulation scheme, we assume that U is distributed normally. How does this method work with other continuous latent variable data generating processes?
- Can we quickly extend this method to nested structural models and marginal structural models for time to event analysis?
- In the continuous setting, we see that low confounding yields the bootstrapped method superior to the naïve one. In the binary logistic setting, the opposite pattern occurs. Why is this?
- Only in the logistic regression setting can we assume that Y causes W. Can we extend this for the continuous setting? If not, why can't we?

Software for the Bayesian Estimator can be found at this [github repository]. In addition, code for reproducing simulation results can be found at this [github repository].

References

Angrist JD, Imbens GW, Rubin DB (1996) Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association 91(434):444–455

Arora N, Papapanou PN, Rosenbaum M, Jacobs DR, Desvarieux M, Demmer RT (2014) Periodontal infection, impaired fasting glucose and impaired glucose tolerance: results from the Continuous National Health and Nutrition Examination Survey 2009-2010. Journal of Clinical Periodontology 41(7):643–652

Christensen RHB (2022) ordinal—Regression Models for Ordinal Data

Dawid AP (2000) Causal Inference Without Counterfactuals. Journal of the American Statistical Association 95(450):407–424, publisher: [American Statistical Association, Taylor & Francis, Ltd.]

Gabry J, Češnovar R, Johnson A, Bronder S (2024) cmdstanr: R Interface to 'CmdStan'

Hoff PD (2009) Latent variable methods for ordinal data. In: Hoff PD (ed) A First Course in Bayesian Statistical Methods, Springer, New York, NY, pp 209–223

Lauritzen SL (1996) Graphical Models. Oxford Statistical Science Series, Oxford University Press, Oxford, New York

Lauritzen SL, Spiegelhalter DJ (1988) Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. Journal of the Royal Statistical Society Series B: Statistical Methodology 50(2):157–194

Liu J, Park C, Li K, Tchetgen EJT (2024) Regression-Based Proximal Causal Inference. Version Number: 3

Montgomery DC (2013) Design and analysis of experiments, eighth edition edn. John Wiley & Sons, Inc, Hoboken, NJ

Pearl J (1988) Probabilistic Reasoning in Intelligent Systems. Elsevier

Pearl J (1993) [Bayesian Analysis in Expert Systems]: Comment: Graphical Models, Causality and Intervention. Statistical Science 8(3):266–269, publisher: Institute of Mathematical Statistics

R Core Team (2023) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Reisach A, Seiler C, Weichwald S (2021) Beware of the Simulated DAG! Causal Discovery Benchmarks May Be Easy to Game. In: Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 34, pp 27772–27784

Robins J (1986) A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling 7(9):1393–1512

Simpson EH (1951) The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society Series B: Statistical Methodology 13(2):238–241

Splawa-Neyman J, Dabrowska DM, Speed TP (1990) On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. Statistical Science 5(4):465–472, publisher: Institute of Mathematical Statistics

Tchetgen EJT, Ying A, Cui Y, Shi X, Miao W (2020) An Introduction to Proximal Causal Learning. ArXiv:2009.10982 [stat]

Tchetgen Tchetgen EJ, Ying A, Cui Y, Shi X, Miao W (2020) An Introduction to Proximal Causal Learning. ArXiv:2009.10982 [stat]

Vander Weele TJ, Shpitser I (2011) A New Criterion for Confounder Selection. Biometrics 67(4):1406–1413

Wright S (1921) Correlation and Causation. Journal of Agricultural Research 20:557–585

Yule GU (1902) Variation of the Number of Sepals in Anemone Nemorosa. Biometrika 1(3):307

A Logistic Regression Results

Number of Samples	Median Squared Error	Simulation Type
100	0.6405	Bayesian Bootstrap
1000	0.3772	Bayesian Bootstrap
100	0.8312	Naive
1000	0.4993	Naive
100	0.2356	Oracle
1000	0.0210	Oracle

Table 6: Error = $(\hat{\beta_T Y} - b_{TY})^2$ Improved point estimate accuracy when using a Bayesian Bootstrap. 10000 simulations for 100 and 1000 sample datasets.

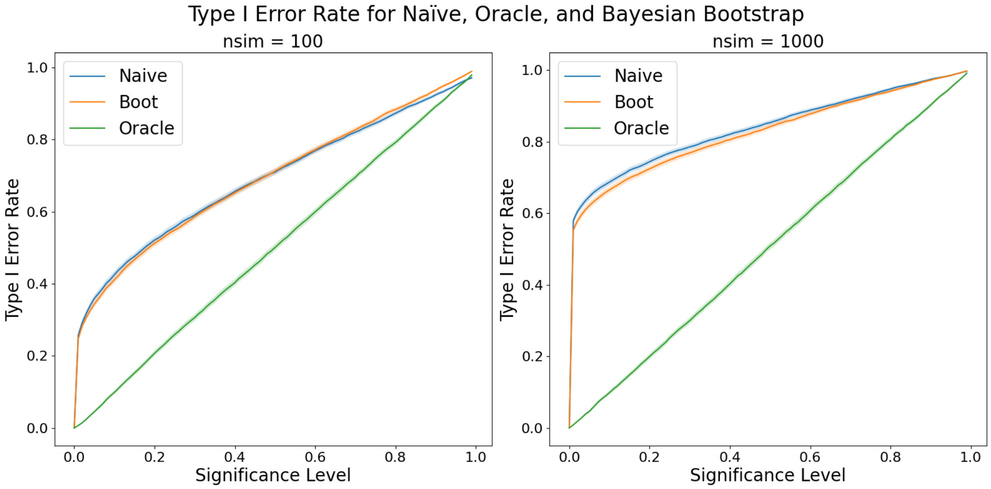


Figure 12: Type 1 error rate comparison with Bootstrap technique performing better than the naive adjustment.

B Prediabetes Reported Estimates

	Periodontal Disease Severity	
Model	Moderate Periodontis	Severe Periodontis
Oral Glucose	1.07 (0.50–2.25)	1.93 (1.18–3.17)
Impaired Fasting Glucose	1.14 (0.74–1.77)	1.12 (0.58–2.18)

Table 7: Reported model coefficients that inspire most of the researchers results. Controlling for socio demographic variables, health behaviors, and adiposity (Body mass index).

	Periodontal Disease Severity	
Model	Moderate Periodontis	Severe Periodontis
Oral Glucose	1.04 (0.52–2.06)	1.75 (1.16–2.62)
Impaired Fasting Glucose	1.14 (0.79–1.65)	1.14 (0.68–1.90)

Table 8: Reported model coefficients for a supplementary model where the researchers only control for socio-demographic variables..