Research Article

Shane Sparkes*, Erika Garcia, and Lu Zhang

The functional average treatment effect

https://doi.org/10.1515/jci-2023-0076
received November 20, 2023; accepted September 27, 2024

Abstract: This article establishes the functional average as an important estimand for causal inference. The significance of the estimand lies in its robustness against traditional issues of confounding. We prove that this robustness holds even when the probability distribution of the outcome, conditional on treatment or some other vector of adjusting variables, differs almost arbitrarily from its counterfactual analogue. This article also examines possible estimators of the functional average, including the sample mid-range, and proposes a new type of bootstrap for robust statistical inference: the Hoeffding bootstrap. After this, the article explores a new class of variables, the $\mathcal{U}$ class, that simplifies the estimation of functional averages. This class of variables is also used to establish mean exchangeability in some cases and to provide the results of elementary statistical procedures, such as linear regression and the analysis of variance, with causal interpretations. Simulation evidence is provided. The methods of this article are also applied to a National Health and Nutrition Survey data set to investigate the causal effect of exercise on the blood pressure of adult smokers.

Keywords: causal inference, functional average, extreme order statistics, mean exchangeability, linear regression, Hoeffding bootstrap

MSC 2020: 60E05, 62J99, 62G30, 62G32

1 Introduction

The Neyman-Rubin (NR) model is an important framework for causal inference with observational designs [1–4]. Say we are interested in studying a population of random variables $\{Y_i\}_{i\in I}$, where $I = \{1, \ldots, N\}$, conditional on exposure to T_i. Here, we specify that each T_i is an indicator variable for simplicity and denote an arbitrary outcome variable Y_i that is observed conditional on $T_i = t$ as $Y_i|t$. At any point in time, it is impossible to observe both $Y_i|1$ and $Y_i|0$ for an arbitrary unit i: a fact that makes individual contrasts, such as $Y_i|1 - Y_i|0$, undefined. This problem also characterizes the "fundamental problem of causal inference" since differences of this type are also undefined in an experimental setup [5]. Since each $Y_i|t$ is observed in the absence of experimental randomization, there is no guarantee that their properties align with the properties of their experimental counterparts: a fact that is problematic for scientific inference. The NR model frames these challenges in the language of missing outcomes and addresses it by constructing populations of counterfactual probability distributions, say $\{Y_i(1)\}_{i\in I}$ and $\{Y_i(0)\}_{i\in I}$. These populations represent a hypothetical situation such that (s.t.) the treatment status of the entire population has been experimentally fixed to a sequence of treatment exposures. The goal under the NR framework is to identify *summary* causal effects that would have occurred, provided one could actually have constructed and manipulated both $\{Y_i(1)\}_{i\in I}$ and $\{Y_i(0)\}_{i\in I}$ [6]. More specifically, under this rubric, for some sample ζ of n outcome variables s.t. $\zeta = \{Y_1|1, \ldots, Y_{n_1}|1\} \cup \{Y_1|0, \ldots, Y_{n_0}|0\}$

*** Corresponding author: Shane Sparkes,** Department of Population and Public Health Sciences, University of Southern California, Los Angeles, California 90033, United States, e-mail: sgugliel@usc.edu
Erika Garcia: Department of Population and Public Health Sciences, University of Southern California, Los Angeles, California 90033, United States, e-mail: garc991@usc.edu
Lu Zhang: Department of Population and Public Health Sciences, University of Southern California, Los Angeles, California 90033, United States, e-mail: lzhang63@usc.edu

and $n = n_1 + n_0$, we wish to identify some multivariate function g and some vector of adjusting variables $\mathbf{L}$ s.t. $E\{g(Y_1|1, \ldots, Y_{n_1}|1) \mid \mathbf{L}\} - E\{g(Y_1|0, \ldots, Y_{n_0}|0) \mid \mathbf{L}\} = E\{g\{Y_1(1), \ldots, Y_N(1)\} \mid \mathbf{L}\} - E\{g\{Y_1(0), \ldots, Y_N(0)\} \mid \mathbf{L}\}$, at least asymptotically [7]. Here, we again use the difference of estimands only as an example. When this is possible, unmeasured variables are said to be ignorable, conditional on $\mathbf{L}$, and the conclusions of the observational study are equivalent to those of an experimental one with randomized treatment exposures [8,9].

Although g can be any function that has scientifically meaningful properties, a small number of summary functions and estimands have dominated the literature. Functions and estimands related to quantiles have commandeered some attention [7,10–12]. A lion's share, however, has been claimed by the arithmetic mean and expected values. Unfortunately, the adage that "there is no free lunch" applies to these common setups. When a researcher wishes to exchange $E\{Y(t)\}$ with an equivalent and identifiable expected value to estimate an *expected* treatment effect, for instance, a set of sufficient conditions are usually required: consistency (C1), strong ignorability (C2), and positivity (C3) [13]. These conditions will soon be defined in detail. For now, it suffices to state that $E(Y|t, \mathbf{L}) = E\{Y(t)|\mathbf{L}\}$ for an appropriately chosen $\mathbf{L}$ under these assumptions. Since the expectation on the left of the equality sign is identified, it can be estimated directly "in place" of the counterfactual expected value on the right – a condition that has typically been labeled as mean exchangeability. Standardization (iterated expectation) is then used to yield the quantity of interest since $E\{E\{Y(t)|\mathbf{L}\}\} = E\{Y(t)\}$. A cardinal problem, however, is that C2 is exceptionally nontrivial to achieve [14–16]. Put succinctly, properly specifying a vector of adjusting variables $\mathbf{L}$ that is sufficient for supplying C2 requires a comprehensive understanding of the causal dynamics involved. However, this level of understanding is absent by design in most circumstances. Consequently, it is reasonable to assert that most researcher-specified $\mathbf{L}$ are likely to fall short of this mark. Finally, $\mathbf{L}$ is often high in dimension. Parametric methods must then be employed to approximate the mean model of Y in conjunction with standardization and bias likely ensues outside of toy examples [6].

One of our cardinal contributions is to highlight a different summary causal effect – the functional average causal effect – as a valuable estimand for the NR framework. While the expected causal effect can detect a change in the weighted average of the observable values of an outcome variable under treatment, the functional average causal effect can capture a change in the uniform average only. This makes the functional average causal effect a coarser estimand that is less flexible but more fundamental than its weighted counterpart. Put otherwise, although it can capture changes in the underlying structure of a random variable, it cannot detect changes in the distribution of these values when the distribution itself is not uniform. Hence, functional average causal effects are relevant for statisticians in research contexts s.t. the treatment or exposure being investigated is expected to transform the set of values that are possible to observe. Examples of such effects are provided in Section 2.

Since functional averages are much more elementary than traditional estimands, however, they are identified under very mild conditions and therefore avoid many of the identification challenges previously discussed. Moreover, we show that they impart a causal interpretation to the results of standard statistical procedures when used in conjunction with a more general type of causal theory that allows for unknown and unmeasured variables. For example, if $Y|t$ and $Y(t)$ have the same image in the traditional analysis sense for all t of interest, this is sufficient for identifying functional average causal effects insofar as the researcher has grounds for defending the assertion that T causally impacts the outcome. The probability distributions of $Y|t$ and $Y(t)$ can otherwise differ arbitrarily. Confounding is immaterial, insofar as it does not change the image of the underlying function(s). So is informative sampling more generally. All that matters for identification is that – theoretically – $Y|t$ and $Y(t)$ pull from the same set of real numbers, at least conditional on some $\mathbf{L}$. Although this is not a necessary condition for what we call functional average exchangeability, it is at least sufficient for establishing it.

The remainder of this article goes as follows. For clarity, we mathematically define functional averages and prove an elementary but fundamental claim in Section 2. Moreover, we show that the functional average treatment effect is salient when the researcher believes that an intervention alters the set of possible values that an outcome can achieve as aforementioned. After this, we examine a small set of functional average estimators, including the sample mid-range, and establish their statistical consistency under general conditions. Since their sampling distributions are largely intractable, we re-purpose the bootstrap as a method for conservative inference. In Section 3, we provide elucidation on a particular class of bounded random variables – the $\mathcal{U}$ class – that generalizes the notion of symmetry and assists in the estimation of functional averages. We also show that $\mathcal{U}$ random variables possess many favorable properties when it comes to causal inference.

These facts allow us to also prove – under the auspices of a causal theory – that linear regressions estimate causal effects under a standard set of assumptions already employed for association studies. Section 4 presents simulation evidence that substantiates our claims.

Finally, in Section 5, we use our strategies in conjunction with data from the National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS) to investigate if a history of exercise activity causally impacts the systolic blood pressure (SBP) of adult smokers. Plentiful evidence exists that smoking is associated with cardiovascular disease processes and mortality [17,18]. Evidence has also been presented that smoking is a factor in arterial stiffening [19]. However, while some literature has supported the proposition that exercise lowers arterial blood pressure [20] and that smokers who exercise show fewer signs of arterial stiffening [21], the evidence is not yet definitive. Functional average estimation targets deterministic changes in the structure of an outcome variable and is thus an informative tool in this context.

2 Functional average treatment effect

In this section, we first introduce important definitions and notation, although some concepts will be left implicit for readability. For instance, we leave the underlying probability space of the form $(\Omega, \mathcal{F}, \mathcal{P})$ for an arbitrary random variable $Y(\omega) : \Omega \to \mathbb{R}$ unstated, and the same goes for probability spaces defining joint distributions. Recall that the support of a random variable is a smallest closed set $\mathcal{S}$ s.t. $\Pr(Y \in \mathcal{S}) = 1$. Alternatively, it can also be defined as the closure of the set of values $\mathcal{S}$ s.t. the density or mass function $f(y) > 0$ for $\forall y \in \mathcal{S}$. Here, we will be dealing with bounded random variables, which means that $\mathcal{S}$ is a strict subset of the real numbers. This is not a limiting constraint. Anything that can be empirically measured is necessarily bounded.

With these concepts, we can revisit the functional average. If $\mathcal{S}$ is discrete, define $R = |\mathcal{S}|$, where $|\cdot|$ in this context denotes the number of elements in the set. If Y is continuous, then $R = \int_{\mathbb{R}} \{1_{y \in \mathcal{S}}\} dy$ and the functional average $Av(\cdot)$ is $Av(Y) = R^{-1} \int_{\mathbb{R}} \{y 1_{y \in \mathcal{S}}\} dy$. For discrete variables, it is $Av(Y) = R^{-1} \sum_{y \in \mathcal{S}} y$. Note that we have avoided the use of general measures for purposes of accessibility.

Sometimes it will be the case that, for some measurable function g, $Y = g(X_1, \ldots, X_k)$. Then the support of Y with respect to (w.r.t.) the joint distribution of $(X_1, \ldots, X_k) = \mathbf{X} \in \mathbb{R}^k$ is some general region $\mathcal{R} \subseteq \mathcal{S}_1 \times \ldots \times \mathcal{S}_k$, where each $\mathcal{S}_i$ indicates the support of X_i. Without loss of generality (WLOG), we will henceforth deal only with the continuous case. In this context, $R = \int_{\mathbb{R}^k} 1_{(x_1, \ldots, x_k) \in \mathcal{R}} dx_1 \ldots dx_k$ and $Av_{\mathbf{x}}(Y) = R^{-1} \int_{\mathcal{R}} g(x_1, \ldots, x_k) dx_1 \ldots dx_k$.

Now, let $E_h Y$ indicate that the expectation of Y is taken w.r.t. a different density or mass function $h(y)$ that is also defined on $\mathcal{S}$. Then it is also apparent that $Av(Y) = E_h Y$ when $h(y)$ is a uniform density or mass function since it then follows that $h(y) = R^{-1}$ and hence $E_h Y = \int_{\mathcal{S}} y \cdot R^{-1} dy = Av(Y)$. When a subscript is omitted and $Y \sim f(y)$, it will be understood that the expectation is taken w.r.t. the baseline density (mass) function f, provided it exists. Otherwise, we say that $E_U Y = Av(Y)$ as a special case, although we will avoid this notation after this section. This is because $Av(Y)$ is best interpreted with the lenses of basic, deterministic analysis. The exception to this statement is when Y truly follows a uniform probability law.

A functional average treatment effect then – for any two treatment values of interest t, t' – can be defined as $h\{Av\{Y(t)\}, Av\{Y(t')\}\} = h\{E_U\{Y(t)\}, E_U\{Y(t')\}\}$ for a user-specified function h. In this article, we set h to a simple difference for exploratory purposes, i.e.,

$$h\{Av\{Y(t)\}, Av\{Y(t')\}\} = \Delta_{t,t'} = Av\{Y(t)\} - Av\{Y(t')\}.$$

2.1 Examples of applicability

The average functional value is not a usual focus in statistical settings. The expected value w.r.t. the baseline measure has instead largely been an object of interest. Hence, we offer a short argument and demonstration of

its importance as a preliminary apologia. We start with the meaning of an expected value. By definition, an expected value is a sum of all possible values that an outcome variable can take, where each value is weighted by the probability of observing it or its density. However, the chance (or density) of observation is extraneous to causal relationships that are unrelated to altered probabilities. Put otherwise, we are not always directly interested in the *chance* of observing an individual outcome under a treatment condition. Mostly, we conceptualize these values to identify summary effects. While these summary effects are interpretable, their meaning is somewhat obscure. For instance, observe $E\{Y(1)\} = \sum_{y \in \mathcal{S}_{Y(1)}} f_{Y(1)}(y) \cdot y$. We often interpret $E\{Y(1)\}$ as the "average value of Y" under experimental treatment. However, this is not entirely accurate since it is a *weighted* average. Language relating to the weighted nature of the sum is often omitted since it does not possess a straightforward empirical interpretation in nontrivial observational contexts. At least partially, this is because the weights are often unknown and cannot be approximated from the data without fairly strong assumptions.

This is not the case for functional averages. Although the functional average can also be construed, albeit counterfactually in most cases, as an expected value, the uniform measure imbues it with a more deterministic interpretation. To appreciate this further, again observe that $Av\{Y(1)\} = |\mathcal{S}_{Y(1)}|^{-1} \sum_{y \in \mathcal{S}_{Y(1)}} y$ in our current context. Hence, in contrast to $E\{Y(1)\}$, it truly is the "average value of Y" under experimental treatment. The unknown probabilities (or densities) do not enter the picture. As a result, functional averages do not require a probabilistic framework to acquire meaning. This fact helps with their interpretation. Moreover, it potentially makes functional averages very salient estimands in the presence of confounding.

We now provide three concrete examples of functional average treatment effects and contrast them with expected treatment effects. The first example instantiates a situation where a functional average effect exists, but an expected one does not. After this, a second example explores the situation where both exist and are nonidentical. Finally, the third example discusses modeling contexts where functional averages are inappropriate estimands in comparison to traditional ones.

2.1.1 Example 1

Say T is a binary treatment variable s.t. $T = 1$ when a particular psychotropic medication is received and Y is a Likert scale measuring anxiety in individuals with clinical depression. Also say that $Y(0)$ can take any integer between one and ten with the following probabilities, respectively:

$$\{0.01, 0.04, 0.05, 0.1, 0.15, 0.15, 0.3, 0.1, 0.05, 0.05\}.$$

Then $E\{Y(0)\} = 6.14$ and $Av\{Y(0)\} = 10^{-1} \sum_{i=1}^{10} i = 5.5$. Now, say $Y(1)$ has nonzero mass only on integers between one and eight with the following probabilities:

$$\{0.01, 0.01, 0.01, 0.05, 0.1, 0.5, 0.18, 0.14\}.$$

Under this scheme, it is also the case that $E\{Y(1)\} = 6.14$. This makes the detection of a causal effect impossible if only the expected treatment effect is utilized. However, $Av\{Y(1)\} = 8^{-1} \sum_{i=1}^{8} i = 4.5$, which means that $\Delta_{1,0} = 4.5 - 5.5 = -1$. This value partially reflects the elimination of the probability that the two most extreme levels of anxiety are observed under treatment, albeit at the cost of a higher probability of mild to moderate anxiety experiences. In other words, after treatment, the respondent will never select nine or ten as measures of their symptom experience, even if provided the opportunity to do so. Changes in functional averages are capable of capturing an aspect of this effect.

2.1.2 Example 2

Again, let T be a binary treatment variable for simplicity and say $Y(0)$ is a random variable for the untreated systolic blood pressure (SBP) of individuals who have been diagnosed with high blood pressure. We will assume that $Y(0)$ follows a truncated normal distribution with a mean at 155 mmHg and support

$\mathcal{S}_{Y(0)} = [110, 370]$ and $Y(1)$ follows a truncated normal distribution with mean 125 mmHg on support $\mathcal{S}_{Y(1)} = [90, 250]$. Here, both an expected and a functional average treatment effect are present and relevant. WLOG, note that $\int_{\mathcal{S}_{Y(0)}} \mathrm{d}y = \int_{110}^{370} \mathrm{d}y = 370 - 110 = 260$. Then $Av\{Y(0)\} = 260^{-1}\int_{110}^{370} y\mathrm{d}y = 240$, while $Av\{Y(1)\} = 170$ by similar calculation. Therefore, $\Delta_{1,0} = 170 - 240 = -70$, while $E\{Y(1)\} - E\{Y(0)\} = -30$. The functional average in this example offers additional information about changes in the possibilities of extremes that cannot be captured by expected values with the baseline probability measure, insofar as it is not uniform.

Pertinently, these examples elucidate how functional averages and their differences remain invariant to any redistribution of the presented probabilities insofar as they remain nonzero on the same set. For instance, say we employed a biased sampling mechanism (such as a convenience sampling) and we also failed to measure the confounders that are sufficient for mean exchangeability in this example. As a consequence, say $Y|1$ follows a truncated normal distribution s.t. $E(Y|1) = 110$ and $Y|0$ follows a truncated normal distribution s.t. $E(Y|0) = 180$, a state of affairs that evidences confounding. This would not matter for our purposes, however, insofar as the convenience sample was executed in such a way as to preserve the sets of values that the functions could theoretically materialize under an experimental design. Under this scenario, although the expected causal effect is confounded, the functional average causal effect remains identified. It is resistant to latent unknowns.

2.1.3 Example 3

As previously stated, the functional average is not a useful estimand in all contexts. For example, in any context s.t. $\mathcal{S}_{Y(1)} = \mathcal{S}_{Y(0)}$, a functional average causal effect will not exist; it will be zero. In fact, this point generalizes to any estimand that is a function of the supports. In a temporary abuse of set notation, this is because if $\mathcal{S}_{Y(1)} = \mathcal{S}_{Y(0)}$, it then follows that $g\{\mathcal{S}_{Y(1)}\} - g\{\mathcal{S}_{Y(0)}\} = 0$ for any function g.

To show this more concretely, again consider the content of Example 2, except say $Y(1)$ – treated SBP – follows a truncated normal distribution with an expected value of 125 mmHg and $Y(0)$ – untreated SBP – follows a truncated normal distribution with an expected value of 180 mmHg. Furthermore, assume that both distributions are now supported on $\mathcal{S}_{Y(0)} = \mathcal{S}_{Y(1)} = [80, 350]$. Note that the behavior of the densities can differ greatly, depending on how we specify the variances. We simply require for any $\varepsilon > 0$ that $f(80 + \varepsilon) > 0$ and $f(350 - \varepsilon) > 0$ WLOG. Consequently, it is immediately observable that $E\{Y(1)\} - E\{Y(0)\} = -55$ and an expected causal effect is present. However, $\Delta_{1,0} = 0$ by construction.

In addition, consider a situation s.t. the outcome is a binary variable and the treatment exposure does *not* eliminate the chance that the event associated with the outcome variable is observed. For instance, state that $Y(1)$ is the occurrence of a mood episode while taking a mood-stabilizing medication and $E\{Y(1)\} > 0$. Under the premise that $Y(0)$ is also nondegenerate, it thus follows that $\mathcal{S}_{Y(1)} = \mathcal{S}_{Y(0)} = \{0, 1\}$ and $\Delta_{1,0} = 0$. An expected causal effect, if identified, would be able to detect if the medication under investigation lowers the probability that a mood episode occurs. Functional averages would not be capable of detecting a change.

Not only do these examples show that the expected and functional average treatment effects are not equivalent when the supports of the experimental outcomes are the same, but they also illustrate that functional average causal effects – and any causal effect predicated upon functions of supports – are inappropriate estimands when it is suspected that the effect of exposure is "weak" in the sense that it only causes perturbations to the distribution of the outcome probabilities. In this sense, functional average effects and causal effects based upon support alterations more generally represent a class of stronger causal effects that communicate the impact of the treatment exposure on the more fundamental structure, or "bones," of the outcome variable. In other words, while expected causal effects attempt to summarily capture changes in what one is likely to observe or experience, functional average causal effects attempt to capture what, on average, is even *possible* to observe or experience.

2.2 Identifying and estimating counterfactual functional averages

Next, we prove some basic statements about functional averages under mild conditions and the rubric of informative sampling. For this, we specify a conditional population of interest $P = \{Y_1|t, \ldots, Y_N|t\}$ s.t. $Y_1|t \sim f(y|T = t)$ WLOG. This setup can be defined with additional conditioning or extended to unconditional circumstances, but this is omitted here for brevity. Additionally, observe a complete-case sample $\zeta \subset P$ and a complementary vector of indicator variables $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)$ s.t. $\delta_i = 1$ if and only if $Y_i|t \in \zeta$. Essentially, each δ_i variable determines if $Y_i|t$ is selected for observation by the researcher. We also assume that $E(\delta_i|y_i, t) > 0$ for all $\forall i$, which implies that $E(\delta_i|t) = \pi_i > 0$ for all i, an assumption that is typically called sampling positivity. It is well-known that an arbitrary $Y_i|t \in \zeta$ does not, in general, follow the distribution of the theoretical population [22–25]. Instead, $Y_i|t \in \zeta$ possesses a weighted density or mass function $f_\delta(y_i|t) = \pi_i^{-1}E(\delta_i|y_i, t)f(y_i|t)$. Utilizing this fact, it is important to note, then, that $E(Y_i|t, \delta_i = 1) = \pi_i^{-1}\sigma_{Y_i|t,E(\delta_i|Y_i,t)} + E(Y_i|t)$ for a fixed t. Here, the notation $\sigma_{Y_i|t,E(\delta_i|Y_i,t)}$ denotes the covariance: $E\{Y_i|t \cdot E(\delta_i|Y_i, t)\} - E(Y_i|t)\pi_i$. To see why this last assertion holds, observe WLOG:

$$\begin{aligned} E(Y_i|t, \delta_i = 1) &= \int_{\mathcal{S}|t} y \cdot \pi_i^{-1}E(\delta_i|y_i, t)f(y_i|t) \cdot \mathrm{d}y \\ &= \int_{\mathcal{S}|t} \{y \cdot \pi_i^{-1}E(\delta_i|y_i, t)\} \cdot f(y_i|t) \cdot \mathrm{d}y \\ &= \pi_i^{-1} \cdot E\{Y_i|t \cdot E(\delta_i|Y_i, t)\} \\ &= \pi_i^{-1}\sigma_{Y_i|t,E(\delta_i|Y_i,t)} + E(Y_i|t). \end{aligned}$$

It is also important to consider a basic sufficient condition for when $f_\delta(y|t)$ and $f(y|t)$ share the same support. Insofar as $E(\delta_i|y_i, t) > 0$ for all y_i, this condition is met. This follows since $f_\delta(y_i|t) = \pi_i^{-1}E(\delta_i|y_i, t)f(y_i|t)$. Therefore, when $E(\delta_i|y_i, t) > 0$ for all $y_i \in \mathcal{S}_{Y_i|t}$, $f_\delta(y_i|t)$ can only be nonzero on precisely the same set of values as $f(y_i|t)$.

For conciseness, we often denote $Y_i|t \in \zeta$ as $Y_{\delta_i}|t$ under the implicit assumption that $\delta_i = 1$. We also use $Y_i|T \in \zeta$ or $Y_{\delta_i}|T$ with the understanding that T is fixed to whatever value it takes for unit $i \in I$. We now specify our short list of assumptions more formally. Altogether, we have three sets. The first set is a re-statement of C1–C3 for clarity. The second set (A1–A4) is mostly constituted by strictly weaker versions of C1–C3; its conditions are sufficient for establishing the preservation of experimental supports and therefore the identifiability of functional average causal effects. The last set, which possesses a single assumption (D1), is important for proving the statistical consistency of estimators under very general conditions of probabilistic dependence. The usual provisos that referenced mathematical objects exist are mostly omitted.

C1: $Y_{\delta_i}|t = Y_{\delta_i}(t)|t$ for all contrasted $t \in \mathcal{S}_{T_i}$ and $Y_i|T_i \in \zeta$ (Consistency).

C2: Let $\mathbf{L}_i$ be a vector of random variables. Then for all contrasted $t \in \mathcal{S}_{T_i}$ and $Y_i|T_i \in \zeta$, it is true that $Y_i(T_i) \perp\!\!\!\perp T_i|\mathbf{L}_i$ (Strong ignorability).

C3: $0 < \Pr(T_i = 1|\mathbf{L}_i) < 1$ for all indexes shared with each $Y_i|T_i \in \zeta$ (Positivity).

Next, we introduce conditions A1–A4, which are more important for this investigation.

A1: Let $\mathbf{L}_i$ be a vector of observable random variables. Then there exists a possibly unknown random vector $\mathbf{U}_i$ s.t. $Y_{\delta_i}|t, \mathbf{l}, \mathbf{u} = Y_{\delta_i}(t)|t, \mathbf{l}, \mathbf{u}$ for all contrasted $t \in \mathcal{S}_{T_i}$, $\mathbf{u} \in \mathbf{S}_{\mathbf{U}_i}$, $\mathbf{l} \in \mathcal{S}_{\mathbf{L}_i}$, and δ_i (Existential consistency).

A2: Let $\mathbf{U}_i$ and $\mathbf{L}_i$ be the same as specified earlier. Then for all contrasted $t \in \mathcal{S}_{T_i}$, it is true that $Y_i(T_i) \perp\!\!\!\perp T_i|\mathbf{L}_i$, $\mathbf{U}_i$ (Existential strong ignorability).

A3: $0 < \Pr(T_i = 1|\mathbf{L}_i, \mathbf{U}_i) < 1$ for all indexes shared with each $Y_i|T_i \in \zeta$ (Existential positivity).

A4: The support of $Y_{\delta_i}|T_i, \mathbf{L}_i$ and $Y_i(T_i)|\mathbf{L}_i$ are the same, i.e., for all $Y_i|T_i \in \zeta$, $\mathcal{S}_{Y_i|T_i,\mathbf{L}_i} \subseteq \mathcal{S}_{Y_i(T_i)|\mathbf{L}_i}$ and $\mathcal{S}_{Y_i(T_i)|\mathbf{L}_i} \subseteq \mathcal{S}_{Y_i|T_i,\mathbf{L}_i}$.

Finally, we now introduce the assumption that is used to establish statistical consistency.

D1: For simplicity, now specify that $I = \{1, 2, 3, \ldots, n\}$ labels the elements of ζ. Let $Y_{\delta_i}|T_i$ be an outcome random variable and say that $\mathcal{L}_n = (I, E_n)$ is an undirected graph with node set I and link set E_n s.t. a link $e_{i,j} \in E_n$ between two nodes $i, j \in I$ is present if and only if $\sigma_{Y_{\delta_i}|T_i,Y_{\delta_j}|T_j} \neq 0$. Then the mean degree of this graph $n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n-1}1_{e_{i,j}\in E_n} = \mu_n = o(n)$, where $1_{e_{i,j}\in E_n} = 0$ when $i = j$ by convention and each indicator variable is nonstochastic. This statement can be generalized to Y_{δ_i} conditional on $\mathbf{L}_i$.

A contrast of each C* to their respective A* for * ∈ {1, 2, 3} is apropos. In general, we can assert that each A* is weaker than its traditional counterpart. For instance, C1 posits that the value for $Y_{\delta_i}(t)$, which is the value that would have been observed for this sampled variable had the entire population's exposure been experimentally fixed to t, is the same value that is conditionally observed for $Y_{\delta_i}|t$. A1 states something that is much weaker; it asserts, provided a researcher has also fixed some $\mathbf{L}_i = \mathbf{l}$, that there merely *exists* some random vector $\mathbf{U}_i$ that *could have been measured* s.t. this same state of affairs holds within the $(t, \mathbf{u}, \mathbf{l})$ stratum of Y_{δ_i}. Importantly, $\mathbf{U}_i$ need not be known or measured. Setting $(\mathbf{U}_i, \mathbf{L}_i)$ to a vector of degenerate variables recovers C1. We also note that A1 could be renamed to "consistency upon correction" since it implies that, even if traditional consistency does not hold, it can be achieved provided the right oracle information. Contrarily, A1 cannot be true when consistency is impossible to achieve.

The comparisons between C2 and A2 and C3 and A3 are analogous in spirit. C2 posits, conditional on some measured $\mathbf{L}_i = \mathbf{l}$, that the counterfactual outcomes are conditionally independent of the probability law governing treatment exposure. A2 does not necessarily posit this. Instead, it says this state of affairs is *possible* to achieve provided oracle information about an additional vector of otherwise latent random variables $\mathbf{U}_i$. Once again, $\mathbf{U}_i$ only needs to exist; the researcher does not need to know or measure it. Similar to the previous assumption, A2 fails when strong ignorability is false for *any* assortment of variables. In lieu of repeating this same sequence of logic, we mostly omit it for C3 and A3. We only add that – although it can seem like A1–A3 are stricter than usual since they require the statements to apply to an unknown set of confounders, which is not the case. Anytime a researcher asserts that they have measured some $\mathbf{X}_i$, for instance, s.t. C2 and C3 are true, this simply means that there is some $(\mathbf{L}_i, \mathbf{U}_i)$ s.t. $\mathbf{X}_i = (\mathbf{L}_i, \mathbf{U}_i)$ and the researcher believes that they have successfully identified and measured these variables. This can also be rephrased to say that $\mathbf{X}_i = \mathbf{L}_i$ and $\mathbf{U}_i$ is degenerate at no loss.

Reiterating the meaning of A4 is useful as a stepping stone for further consideration. Put succinctly, when A4 holds, it means that the counterfactual distribution and conditional distribution possess the same support, conditional on some $\mathbf{L}_i = \mathbf{l}$. Rejecting this notion is equivalent to positing that certain values in the support of $Y_i(T_i)|\mathbf{L}_i$ can never materialize with $Y_i|T_i, \mathbf{L}_i$. This is a strong assertion with nontrivial epistemic consequences, especially in the context of noninformative sampling. If it is believed that A4 cannot be obtained, then those values that exist in the counterfactual support alone have no real-world meaning. In this circumstance, we can simply condition on those that can materialize at no empirical loss. It is also important to state that, although A1–A3 are important for reasons soon revealed, A4 alone is sufficient for identifying functional average effects and causal effects based upon supports more generally. Since A1 and A2 might not be necessary conditions for A4, the latter is important as a stand-alone condition. For instance, a researcher can assert A4 when they can feasibly defend their sampling process.

Next, we informally establish that A1–A3 imply A4. This is important since it proves that if a researcher thinks it is even possible to try to identify and estimate expected causal effects via the traditional set of assumptions (C1–C3), the experimental supports are already preserved and all causal estimands based upon them are identified. For demonstrative purposes, we prove this informally for the case s.t. U is an absolutely continuous variable, also at no loss, and that all referenced densities exist. To this end, observe the following identity under the stated premises:

$$
\begin{aligned}
f\{y(t)|l\} &= \int_{\mathcal{S}_U} f^{-1}(l) \cdot f\{y(t), l, u\}\mathrm{d}u \\
&= \int_{\mathcal{S}_U} f^{-1}(l) \cdot f\{y(t)|l, u\}f(l, u)\mathrm{d}u \\
&= \int_{\mathcal{S}_U} f\{y(t)|t, l, u\}f(u|l)\mathrm{d}u \qquad \text{(A2 and A3)} \\
&= \int_{\mathcal{S}_U} f(y|t, l, u)f(u|l)\mathrm{d}u. \qquad \text{(A1)}
\end{aligned}
$$

Now, note that the following statement is also true by similar strokes of logic:

$f(y|t, l) = \int_{\mathcal{S}_U} f(y|t, l, u)f(u|t, l)\mathrm{d}u$. This also follows by A2 and A3.

Since both $f\{y(t)|l\}$ and $f(y|t, l)$ are functions of y alone and otherwise share in $f(y|t, l, u)$ as a basis, it is then implied that $f\{y(t)|l\}$ and $f(y|t, l)$ are strictly positive on the same set of values. Furthermore, since $E(\delta|y, t, l) > 0$ implies that $f_\delta(y|t, l) > 0$ on the same set of y values s.t. $f(y|t, l) > 0$, by transitivity, $f_\delta(y|t, l)$ also shares the same support with $f\{y(t)|l\}$. This supplies A4 and corroborates our assertions.

We conclude the discussion of A4 with one additional point. Namely, we explicitly draw attention to the fact that A4 implies at least a restricted form of positivity when the necessary conditional probabilities are posited to exist. This follows because $f(y|t, l, u)$ WLOG requires that $f(t, l, u) > 0$ by construction, which implies that $f(t|l, u) > 0$. When T is a binary variable, this also necessitates that $f(t'|l, u) < 1$. Pertinently, however, A4 does not need positivity to hold for all possible strata of $\mathbf{L}_i$; it only needs to hold for observed strata. This fact becomes useful in later sections. Finally, we conjecture that A4 does not imply A1 and A2 generally, although no further investigation on this matter is pursued here.

Before discussing D1, we offer one more pivotal exploration, especially since it is a segue into the utility and attitude accompanying our approach. In short, we formulate $Y_i(T_i)|\mathbf{L}_i$ and $Y_{\delta_i}|T_i, \mathbf{L}_i$ in the context of test theory, i.e., we consider a model s.t. the conditional observational outcome is equal to the conditional experimental outcome plus possibly nonindependent error. We do this because we also wish to point out that functional average causal effects are identified outside the paradigms built upon C1–C3 or even A4.

Formalizing this entails the following statement: $Y_{\delta_i}|T_i, \mathbf{L}_i = Y_i(T_i)|\mathbf{L}_i + \varepsilon_i$. Mathematically, this statement is well defined since the observational and experimental outcomes share a probability space. This model posits that the observational outcome, conditional on some vector of scientifically salient variables, is equal to the measurement of the experimental outcome plus biasing noise that results from random measurement error or lurking variables. It is also apropos to state that, unless $\varepsilon_i = 0$ almost surely, consistency is violated by this construction under C2. However, if we additionally assert that, although possibly dependent, $Y_i(T_i)|\mathbf{L}_i$ and ε_i are supported on $\mathcal{S}_{Y_i(T_i)|\mathbf{L}_i} \times \mathcal{S}_{\varepsilon_i}$ s.t. $Av(\varepsilon_i) = 0$, it then follows that $Av\{Y_{\delta_i}|T_i, \mathbf{L}_i\} = Av\{Y_i(T_i)|\mathbf{L}_i\}$. The stipulation that $Av(\varepsilon_i) = 0$ essentially means that the two distributions possess the same values on average, provided the right $\mathbf{L}_i$. Again, this should not be misconstrued as a sufficient condition for mean exchangeability since it is possible that $E\varepsilon_i \neq 0$ within this setup.

We also introduce this conceptualization, albeit briefly, because it has good potential for development and can provide an intuitive medium for causal inference in the presence of pervasive interference and model inaccuracy. Finally, it can also be interpreted using the framework of structural causal theory. If $\mathbf{L}_i$ is a sufficient set for adjustment for the "true" but possibly unknown causal graph, then it does follow that $\varepsilon_i = 0$ almost surely. Hence, one can feasibly surmise that the aforementioned conditions and functional average exchangeability will hold when one's working causal model is incorrect and $\mathbf{L}_i$ is insufficient for adjustment; however, the conceptualization is "close enough" to render the two outcome variables the same on average nevertheless.

Although we do not develop this route in any major way within this article, appreciating this type of modeling still contextualizes our main approach and allows us to pay homage to a fundamental caveat. We do the latter first. Identifying a parameter statistically is insufficient for the provision of causal meaning. Causal relationships cannot be inferred from statistical relationships alone [26]. A theory of causation – perhaps represented by a structural causal model – is therefore still required if causal meanings are to be supplied to the functional average or other estimands based upon supports [2,6]. The pivotal difference with our approach is that it is relatively forgiving of incomplete or inaccurate theories of causation. As previously stated, while other approaches require the correct specification of some random vector $\mathbf{L}$ that is sufficient for adjustment or C2, our approach only requires that such an adjustment is theoretically possible. Estimands based upon supports are identified, for example, with a graph constituted by the node set $\{T, Y, U\}$ and the following links, insofar as the researcher is ready to defend the causal assertions implicit within them: $T \to Y$, $U \to T$, $U \to Y$. In other words, if A4 holds, then, provided a more general structural causal model that allows for unmeasured confounders, the functional average effect, and others, are identified and estimable from data without accounting for the latent forces empirically. In a similar vein to the test error setup introduced in the previous paragraph, which only feasibly requires us to "get close enough" to the experimental outcome, this main approach only requires that we can even try to get close at all.

Finally, we turn to the last assumption, D1. Put succinctly, its purpose is to establish statistical consistency for estimators of potential mean outcomes and functional average outcomes in addition. Results are often proven under the premises of mutual independence and noninformative sampling. This restricts their utility, especially since many modern research settings depend upon nonprobability samples of outcome variables that partake in complicated and unknown systems of possibly "long-range" probabilistic dependence. Furthermore, this is also restricting since informative sampling can induce statistical dependencies. By proving our results under more general conditions, we expand their reliability into these contexts. D1 essentially asserts that the mean number of outcome variables in a sample that a typical one is correlated with is sublinear in n, i.e., that $n^{-1}\mu_n \to 0$ as $n \to \infty$. Note that this is a very mild assumption since it still allows for the mean number of statistical dependencies present in the sample to diverge with sample size. It places no additional constraint on the exact form of the probabilistic dependencies. We also reference an alternative: D1′. This assumption is exactly the same, except it makes use of a dependence graph s.t. a link exists between two nodes if and only if their corresponding outcome variables are statistically dependent.

A proof of our first proposition, which formally establishes functional average exchangeability via A4, concludes this section. Although it is trivial mathematically, it provides a useful foundation.

Proposition 1. (Functional average exchangeability) *Suppose A4. Then* $Av(Y_\delta|T, \mathbf{L}) = Av\{Y(T)|\mathbf{L}\}$.

Proof. The result follows directly from the premises. Since $\mathcal{S}_{Y_\delta|T,\mathbf{L}} = \mathcal{S}_{Y(T)|\mathbf{L}}$, $\int_{\mathcal{S}_{Y_\delta|T,\mathbf{L}}} 1 \cdot \mathrm{d}y = \int_{\mathcal{S}_{Y(T)|\mathbf{L}}} 1 \cdot \mathrm{d}y = R$ WLOG for the continuous case and $\int_{\mathcal{S}_{Y_\delta|T,\mathbf{L}}} y\mathrm{d}y = \int_{\mathcal{S}_{Y(T)|\mathbf{L}}} y\mathrm{d}y$. □

Proposition 1 is extendable to $Av_1\{Y(T)|\mathbf{L}\}$. However, again, this is omitted. From here, the notation for $\mathbf{L}$ is sometimes omitted for brevity. We do the same, when possible, for sample indexes.

2.3 The problem of estimation

The simplicity of Proposition 1 and the relative mildness of A1–A3 unfortunately coexist with the difficulty of estimating $Av(Y_\delta|T)$. Here, the "there is no free lunch" adage returns. A theoretical estimator can be constructed, nevertheless, using the following two identities. We tacitly condition on $1_{\mathcal{S}_Y}$ for ease of reading: $E\{f^{-1}(Y)Y\} = \int_{\mathcal{S}} y\mathrm{d}y$ and $E\{f^{-1}(Y)\} = R$. This naturally suggests an estimator of the following form:

$$\tilde{Av}(Y_\delta) = \left\{\sum_{i=1}^{n} f_{\delta_i}^{-1}(Y_{\delta_i})\right\}^{-1} \sum_{i=1}^{n} f_{\delta_i}^{-1}(Y_{\delta_i})Y_{\delta_i}. \tag{2.1}$$

In this section, we investigate some of the features of plug-in estimators for equation (2.1). After exploring the discrete case, we offer brief commentary on the difficulties of the continuous one. Then we re-visit the sample mid-range estimator. After completing these explorations, we introduce a bootstrapping strategy for conducting inference. Pertinently, we also explore these estimators and the challenges surrounding them to foster more appreciation for the utility of an alternative route that is proposed in Section 3.

2.3.1 Discrete estimators of $\tilde{Av}(Y_\delta)$

When Y_{δ_i} is discrete and $\hat{f}_n(y) = n^{-1}\sum_i^n 1_{Y_{\delta_i}=y} = n^{-1}M_y$, the empirical plug-in for equation (2.1) reduces to an intuitive estimator. Say $1_{y\in\mathcal{S}_\zeta}$ is an indicator that a value $y \in \mathcal{S}_Y$ is observed and therefore in the support of the empirical distribution: $\mathcal{S}_\zeta$. Then the empirical plug-in for equation (2.1) reduces to $\hat{Av}(Y_\delta) = \{\sum_{y\in\mathcal{S}_Y} 1_{y\in\mathcal{S}_\zeta}\}^{-1}\sum_{y\in\mathcal{S}_Y} 1_{y\in\mathcal{S}_\zeta} y$ under the convention that $\hat{f}_n^{-1}(y) = 0$ when $y \notin \mathcal{S}_\zeta$. To see this, observe that for an arbitrary set of materialized sample values in ζ, where ζ is temporarily treated as a set of constants, $\sum_{y\in\zeta}\hat{f}_n^{-1}(y) = n \cdot |\mathcal{S}_\zeta|$ and $\sum_{y\in\zeta}\hat{f}_n^{-1}(y)y = n \cdot \sum_{y\in\mathcal{S}_\zeta} y$.

Therefore, the discrete plug-in for equation (2.1) is simply the arithmetic average of the unique values observed in the sample.

We now establish the statistical consistency of this plug-in under general dependency conditions. For the rest of this section, we omit notation for δ with the understanding that it is implicit whenever we are dealing with sampled outcomes. To this end, note that $\Pr(Y_1 \neq y, Y_2 \neq y, \ldots, Y_n \neq y) = \Pr(Y_n \neq y | Y_{n-1} \neq y, \ldots, Y_1 \neq y) \cdot \Pr(Y_{n-1} \neq y | Y_{n-2} \neq y, \ldots, Y_1 \neq y) \ldots \Pr(Y_1 \neq y)$ and define a corresponding sequence $\mathcal{F} = (\Pr(Y_i \neq y | \mathcal{A}_i))_{i \in I}$ under the convention that $\Pr(Y_1 \neq y | \mathcal{A}_1) = \Pr(Y_1 \neq y)$.

Proposition 2. *Suppose a sample $\zeta = \{Y_i\}_{i \in I}$. Observe $\mathcal{F} = (\Pr(Y_i \neq y | \mathcal{A}_i))_{i \in I}$ as previously defined and say $k(n) = |\{s \in \mathcal{F} | s < 1\}|$, where $s \in \mathcal{F}$ indicates here that s is present in the sequence. If $k(n) \to \infty$ as $n \to \infty$, then $\hat{Av}(Y) \xrightarrow{a.s.} Av(Y)$ as $n \to \infty$, where $\xrightarrow{a.s.}$ denotes almost sure convergence.*

Proof. Let $y \in \mathcal{S}_Y$ be arbitrary and denote $\mathcal{S}_\zeta \subseteq \mathcal{S}_Y$ as the set of observed values. Then $\Pr(y \in \mathcal{S}_\zeta) = 1 - \Pr(y \notin \mathcal{S}_\zeta) = 1 - \Pr(Y_1 \neq y, Y_2 \neq y, \ldots, Y_n \neq y) = 1 - \Pr(Y_n \neq y | Y_{n-1} \neq y, \ldots, Y_1 \neq y) \cdot \Pr(Y_{n-1} \neq y | Y_{n-2} \neq y, \ldots, Y_1 \neq y) \ldots \Pr(Y_1 \neq y)$. Now, suppose $k(n)$ probabilities in the sequence $\mathcal{F}$ are strictly less than one. Denote the maximum of these probabilities as $\Pr(Y_* \neq y)$ and note that since $\Pr(Y_* \neq y) < 1$, there exists some $\varepsilon > 0$ s.t. $\Pr(Y_* \neq y) = 1 - \varepsilon$. Then:

$$\begin{aligned}\Pr(y \notin \mathcal{S}_\zeta) &= \Pr(Y_n \neq y | Y_{n-1} \neq y, \ldots, Y_1 \neq y) \cdot \Pr(Y_{n-1} \neq y | Y_{n-2} \neq y, \ldots, Y_1 \neq y) \ldots \Pr(Y_1 \neq y) \\ &\leq 1^{n-k(n)} \cdot \{\Pr(Y_* \neq y)\}^{k(n)} \\ &= \{1 - \varepsilon\}^{k(n)}.\end{aligned}$$

Hence:

$$0 \leq \lim_{n \to \infty} \Pr(y \notin \mathcal{S}_\zeta) \leq \lim_{n \to \infty} \{1 - \varepsilon\}^{k(n)} = 0.$$

This of course implies that $\Pr(y \in \mathcal{S}_\zeta) \to 1$ as $n \to \infty$. Next, define an indicator variable $1_{y \in \mathcal{S}_\zeta}$ and also $Z_{k_*} = \sup_{k > n} |1_{y \in \mathcal{S}_{\zeta_k}} - \Pr(y \in \mathcal{S}_{\zeta_k})| = |1_{y \in \mathcal{S}_{\zeta_{k_*}}} - \Pr(y \in \mathcal{S}_{\zeta_{k_*}})|$. Letting $\varepsilon > 0$ be arbitrary again:

$$\Pr(Z_{k_*} > \varepsilon) \leq \varepsilon^{-2} \{\Pr(y \in \mathcal{S}_{\zeta_{k_*}}) \cdot (1 - \Pr\{y \in \mathcal{S}_{\zeta_{k_*}}\})\}.$$

This then implies that:

$$\lim_{n \to \infty} \Pr(Z_{k_*} > \varepsilon) \leq \lim_{n \to \infty} \varepsilon^{-2} \{\Pr(y \in \mathcal{S}_{\zeta_{k_*}}) \cdot (1 - \Pr\{y \in \mathcal{S}_{\zeta_{k_*}}\})\} = 0.$$

Hence, $1_{y \in \mathcal{S}_\zeta} \xrightarrow{a.s.} 1$. Thereby, since y was arbitrary, it is then implied that $\hat{Av}(Y) = \{\sum_{y \in \mathcal{S}_Y} 1_{y \in \mathcal{S}_\zeta}\}^{-1} \sum_{y \in \mathcal{S}_Y} 1_{y \in \mathcal{S}_\zeta} y \xrightarrow{a.s.} R^{-1} \sum_{y \in \mathcal{S}_Y} y = Av(Y)$ since R is finite. □

The elementary nature of $\hat{Av}(Y)$ makes it a reliable estimator for relatively simple outcome variables. $\hat{Av}(Y)$ will converge almost surely at an unknown, but very fast rate in all likelihood, and even in the presence of stark probabilistic dependencies, when the scale of the outcome variable possesses a small number of unique values. This statement obviously applies to sample extremes in addition.

Unfortunately, however, quantifying the rate of convergence – or the uncertainty associated with finite sample estimates – is difficult. This is true even under mutual independence. To appreciate this, it is sufficient to observe $\sum_{y \in \mathcal{S}_Y} 1_{y \in \mathcal{S}_\zeta} y$. Since $E 1_{y \in \mathcal{S}_\zeta} = p_{y,n}$ is unknown, so is $\mathrm{Var}\{\sum_{y \in \mathcal{S}_Y} 1_{y \in \mathcal{S}_\zeta} y\} = \sum_{y \in \mathcal{S}_Y} p_{y,n} \cdot (1 - p_{y,n}) \cdot y^2$. If ζ is a sample of identically distributed and mutually independent outcome variables, then we can attempt to estimate $p_{y,n}$ with $\hat{p}_{y,n} = 1 - \{1 - \hat{f}_n(y)\}^n$. However, $\hat{p}_{y,n} \equiv 0$ when $y \notin \mathcal{S}_\zeta$. Furthermore, when $y \notin \mathcal{S}_\zeta$, it is also unknown by definition. Hence, reasonably estimating $\mathrm{Var}\{\sum_{y \in \mathcal{S}_Y} 1_{y \in \mathcal{S}_\zeta} y\}$ requires knowledge that makes estimating $Av(Y)$ arguably redundant.

A recourse to the central limit theorem is also unavailable. This is because $|\mathcal{S}_Y|$ is finite by construction. Therefore, $\hat{Av}(Y)$ will always be a finite sum of random variables. In some circumstances – such as when $|\mathcal{S}_Y|$ is reasonably large – a normal approximation might still function with an acceptable degree of accuracy.

However, for reasons already explored, this strategy will still require a strong set of assumptions about sampling probabilities and potential values.

2.3.2 Continuous outcomes

Kernel density estimation is an intuitive choice to estimate equation (2.1) for the continuous case. However, the properties of this plug-in are also largely intractable and unknown. For example, although the properties of a kernel density estimator $\hat{f}_n(y)$ are well researched for a constant $y \in \mathcal{S}_Y$ [27–29], the behavior of $\hat{f}_n(Y)$, i.e., the random variable defined and evaluated on the same random outcome that was utilized to construct it, is not as well studied. This is because kernel density estimation is often evaluated on a grid of deterministic points. Establishing the asymptotic properties of a statistic of the form $\{\sum_{i=1}^n \hat{f}_n^{-1}(Y_i)\}^{-1}\sum_{i=1}^n\{\hat{f}_n^{-1}(Y_i)\}^{-1}Y_i$, where $\hat{f}_n^{-1}(Y_i) = nh \cdot \{\sum_{j=1}^n K\{h^{-1}(Y_i - Y_j)\}\}^{-1}$ for some $h > 0$ and kernel function $K(\cdot)$, although promising, is therefore also nontrivial. Such considerations also require a detailed consideration of possible kernel functions. Since – in general – we are interested in establishing statistical consistency under very general dependency conditions, we avoid this enterprise in this article.

We also avoid other options for density estimation since they arrive with similar challenges, some as yet undisclosed. For instance, estimators that use reciprocal estimated densities can possess unstable variances when the underlying distribution possesses a density that decays smoothly toward zero. Moreover, since each $\hat{f}_n(y)$ is typically a function of the entire sample, plugins for equation (2.1) will necessarily possess a myriad of complex dependencies. This will prevent any elementary citation of a central limit theorem. Just as importantly, it will also limit the applicability of concentration inequalities for finite sample inference. Hence, although this is a promising area of research that demands attention, no further consideration is offered here. Instead, we focus on other estimators for the functional averages of continuous outcomes that do not rely on density estimation.

2.3.3 Mid-range estimation

For a large special class of bounded random variables, the mid-range is a simple alternative for estimating functional averages, including those from continuous distributions. Let $Y_{(i)}$ for $i \in I_n = \{1, \ldots, n\}$ denote the ith order statistic of a sample s.t. $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$. The mid-range, $\hat{MR}\{Y\}$, or simply $\hat{MR}$ when convenient, is defined as follows: $\hat{MR}\{Y\} = 2^{-1}\{Y_{(1)} + Y_{(n)}\}$.

Naturally, the sample mid-range estimates the population mid-range $MR = 2^{-1}(m + M)$. Linear combinations of order statistics are also well studied [30–33]. However, the sample mid-range is often ignored, and especially in applied settings, because of its possible inefficiency and since its distribution also admits no closed-form expression in a majority of situations. Before offering an exposition on some of its properties, we offer a useful definition, which highlights our interest in it. We say a random variable is *regular* when it is supported on a single interval of real numbers or a complete subset of integers. This definition is helpful because $Av(Y) = MR\{Y\}$ when Y is a regular random variable.

Definition 1. A random variable will be said to be regular if and only if its support $\mathcal{S}$ is a single interval of real numbers or a complete subset of integers starting at some $m \in \mathbb{N}$ and ending with a maximum integer M s.t. if integer $c \in \mathcal{S} \backslash M$, then $c + 1 \in \mathcal{S}$.

$\hat{MR}$ is a statistically consistent estimator of MR for outcome variables with finite support under the assumption of mutual independence. Barndorff-Nielsen [34] established sufficient and necessary conditions for the statistical consistency of extreme order statistics. Almost sure convergence, and therefore also convergence in probability ($\xrightarrow{p}$), of an extreme order statistic to its asymptotic target is trivially fulfilled when there

exists a $y \in \mathcal{S}_Y$ s.t. $F(y) = 1$ and $F(y - \varepsilon) < 1$ for all $\varepsilon > 0$. Hence, $\hat{MR}$ also converges almost surely to its population value for all bounded distributions under this setup. Sparkes and Zhang [35] extended this result to much more general scenarios of statistical dependence. If we define a sequence of conditional cumulative distribution functions (CDFs) $\mathcal{F}$ in the same spirit as Proposition 2, it can be demonstrated that extreme order statistics converge in probability to their target values for bounded random variables insofar as the number of conditional CDFs in $\mathcal{F}$ that is strictly less than unity diverges as n becomes arbitrarily large. This is once again a very mild assumption since the dependencies involved can induce arbitrary changes in the behaviors of the distribution functions otherwise. Since the random variables considered here are bounded, convergence in probability of the sample extremes also implies their almost sure convergence.

Nevertheless, as previously mentioned, when the distribution of Y is unknown, no reliable expression for the distribution of $\hat{MR}$ is accessible to use for inference [36–39]: a situation that is analogous to the discrete plug-in estimator for equation (2.1). Additional discussion on this topic is available in the supplementary material. Bootstrapping is a feasible option for inference, provided these challenges. However, it is also not without problems. Traditional bootstraps condition on the observed values of ζ and use the strong consistency of the empirical CDF $\hat{F}_n(y)$ to emulate the sampling distribution of a statistic of interest via a re-sampling procedure [40]. They require the estimand of interest, which is estimated by the statistic $T_0(Y_1, Y_2, \ldots, Y_n)$, to be a well-behaved functional of the marginal CDF. They also stipulate that the targeted parameter is not a boundary value of the support [41]. Overall, bootstrapping processes usually behave as intended under the same set of conditions that supply a central limit theorem. For these reasons, traditional bootstraps are problematic for functions of extreme order statistics. The m-out-of-n bootstrap, however, has proven to be an effective procedure in this domain [42]. Essentially, a basic m-out-of-n bootstrapping process re-samples m observations from ζ with or without replacement s.t. $n^{-1}m \to 0$ as $m \to \infty$. It can provide approximately valid inference when traditional methods fail. See Bickel and Ren [43], Swanepoel [44], Beran and Ducharme [45], or Politis et al. [46] for additional background and resources on the topic. Pertinently, the m-out-of-n bootstrap is also capable of handling situations with dependent observations insofar as an appropriate sub-sampling strategy is used.

Nevertheless, the m-out-of-n bootstrap (and other forms of bootstraps for dependent observations) is still insufficient for the context and conditions of this article. Three reasons substantiate this claim. First, we require a version of the bootstrap that is capable of reliably capturing θ under fairly general but unknowable dependency conditions. This rules out approaches such as the m-out-of-n bootstrap, or bootstrapping processes such as the block bootstrap, which require a re-sampling theory that corresponds adequately to the unknown dependency structure, and which typically exclude the existence of long-range dependencies [47–50].

Second, we are interested in reasoning about $\theta = Av(Y)$ and not $E\{\hat{MR}\}$ for a particular n. In most circumstances where $\hat{MR}$ will be used, i.e., those circumstances s.t. the marginal distributions are not symmetric, it will be a biased estimator [39]. It is likely that θ rests on or near the boundary of the support of $\hat{MR}$ in these circumstances. Convergence to θ might be slow and characterized by an unknown rate in addition [38]. Consequently, any inferential procedure for $\hat{MR}$ must be flexible enough to provide cogent statements about θ – and not simply about $E\{\hat{MR}\}$ at a particular value of n – and even when sample sizes are modest. This necessitates conservative approaches for inference that allow for θ to sit in the extremes of the empirical distribution of the bootstrapped statistics. It also therefore rules out popular bootstrapping methodologies, which construct confidence sets that are strict subsets of the hull of the observed range. Finally, we wish to use a bootstrapping strategy that does not rely on the assumption that θ is a smooth functional of $F(y)$.

Further work to produce more efficient closed-form approximations is of course a preferable route. Since the mid-range is more efficient than the sample mean when nonnegligible probability rests in the extremes of the support [51], it can provide a more efficient estimator of expected causal effects in many circumstances: a fact that is often neglected. Overall, however, due to the complicated probabilistic character of order statistics, this is an onerous road that possesses no immediate destination, especially when outcome variables are dependent and their joint distribution is unknown. This ultimately necessitates a different type of bootstrapping strategy.

2.3.4 The Hoeffding bootstrap

With these prior facts in mind, we offer a novel bootstrapping solution that can address these challenges. In summary, we assert that the bootstrap can be re-purposed to construct conservative confidence sets under fairly general conditions of statistical dependence. Notably, this re-purposed bootstrap, which we call the Hoeffding bootstrap, can be applied to all functional average estimators previously explored since it does not require θ to be a well-behaved functional of $F(y)$.

We now provide a synopsis of the approach. Further details and proofs of its validity and robustness are provided in the supplementary materials. Essentially, we show that (1) if a statistician does not condition on the observed values of ζ and treats each re-sampled Y_i as random, (2) if the outcome variables being re-sampled are not monotonic transformations of one another, say, or they do not partake in other forms of truly extreme statistical dependence, and (3) it is subsequently true that $T_0 \xrightarrow{a.s.} ET_0$ and has finite support, then confidence sets of the form $T_0 \pm \phi(\hat{M}_B - \hat{m}_B)$ provide at least $1 - \alpha$ coverage asymptotically for θ provided any $\phi \geq 1$. $\hat{M}_B$ and $\hat{m}_B$ represent the maximum and minimum order statistics of the bootstrap sample, respectively. Ultimately, we call this method the Hoeffding bootstrap because Hoeffding's inequality is used to specify ϕ. The performance of confidence sets constructed with this strategy are evaluated in Section 4 and also in the supplementary materials. For clarity, we provide a schematic of the process:

I. Acquire a sample of random variables $\zeta = \{Y_i\}_{i \in I_n}$ and compute a statistic $T_0(Y_1, \ldots, Y_n)$
II. Draw $m \leq n$ random variables from ζ with or without replacement via a simple random sample or a theoretically guided process that attempts to reproduce a dependency structure. Compute the new statistic T_1 from these variables
III. Repeat I. and II. $K(n) - 1$ times, where $K(n) = K$ is reasonably large, and construct $\{T_k\}_{k \in \mathcal{K}}$ for $\mathcal{K} = \{0, 1, 2, \ldots, K\}$
IV. Set $\hat{M}_B - \hat{m}_B = \max_{k \in \mathcal{K}}(T_k) - \min_{k \in \mathcal{K}}(T_k)$
V. Construct an estimate of an at least $1 - \alpha$ confidence set with $T_0 \pm \{\hat{M}_B - \hat{m}_B\}\sqrt{2^{-1}\log(2/\alpha)}$

Since gaining some intuition about how this method works is still important at this point, we offer some basic insights. Note that because $\hat{m}_B \leq T_0 \leq \hat{M}_B$ by construction, it is already implied that $ET_0 \in [E(\hat{m}_B), E(\hat{M}_B)]$ with probability one and hence that, for some sufficiently large n and K, $ET_0 \in [\hat{m}_B, \hat{M}_B]$ with probability that is approximately one under very mild regulations. However, since we actually want to capture θ and, in practice, we often use only moderately sized K with moderate n, we require some form of intuitive penalization to mitigate small-sample bias. The use of Hoeffding's inequality, although arbitrary, remains intuitive since its inversion results in the confidence set that would have resulted if one naively used the bootstrapped extremes to estimate the extremes of the support of T_0. In summary, insofar as $T_0 \xrightarrow{a.s.} \theta$ and $\alpha \leq 0.27$ (this ensures that $\phi \geq 1$), employing Hoeffding's inequality to construct a penalty aligns with traditional statistical instincts while guaranteeing more robust small-sample coverage for θ.

Hoeffding's bootstrap also has other benefits. First and foremost, although conservative, it avoids dependency modeling. Since the dependency structure of the outcome variables is invariably unknown, this is a boon. Secondly, as previously stated, it applies to a much larger class of statistics because it does not require θ to be a "smooth" functional of $F(y)$. Moreover, Hoeffding's bootstrap does not require the user to choose a re-sample size like the m-out-of-n bootstrap. As we see in Section 4, the Hoeffding bootstrap can even be more efficient than the m-out-of-n bootstrap for some resampling specifications. Perhaps more importantly, however, is the following statement: that even if the conditions that validate this approach are not met, the Hoeffding bootstrap will always perform better than strategies that use bootstrapped t-statistics or normal approximations. This fact essentially flows from Popoviciu's inequality, which states that $\mathrm{Var}(T_0) \leq 4^{-1}(M_0 - m_0)^2$ when T_0 is bounded. This inequality also applies to empirical distributions. For instance, say that $\alpha = 0.05$. Then $1.96 \cdot S_{T_k} \leq (\hat{M}_B - \hat{m}_B) \leq (\hat{M}_B - \hat{m}_B) \cdot \sqrt{2^{-1}\log(2/\alpha)} \approx 1.35 \cdot (\hat{M}_B - \hat{m}_B)$, where S_k is the sample standard deviation of the bootstrap distribution. Therefore, even in worst-case scenarios, the Hoeffding bootstrap will always provide more robust and cogent statements about uncertainty than a large

proportion of popular methodologies. By the same stroke, it will also provide more conservative coverage. Again, examples of its coverage are provided in Section 4 and the supplementary material.

3 $\mathcal{U}$ random variables and counterfactual linear regression

In this section, we discuss a class of random variables – the $\mathcal{U}$ class – that can help us avoid the difficulties encountered in Section 2. Recall: although we established that functional average causal effects can be identified and statistically consistently estimated under very mild assumptions – and even without adjusting for confounders – establishing efficient methods of statistical inference for these estimators is challenging. Ultimately, this is because the plug-in estimators defined and the sample mid-range possess largely intractable properties, and even asymptotically, in the absence of additional constraints that are in all likelihood inappropriate for applied settings. These difficulties are removed when working with $\mathcal{U}$ random variables as outcomes since they ultimately allow for the functional average to be estimated by standard additive statistics with well-known properties. We also show that $\mathcal{U}$ random variables are important because they can imbue basic linear regressions and analyses of variance with counterfactual – and thus possibly causal – interpretations under conditions traditionally assumed for estimating associations. On a similar note, we also prove that properties of $\mathcal{U}$ variables can be used to establish sufficient conditions for mean exchangeability and that they can be used to defend an extension of the Hoeffding bootstrap. First, however, a definition of a $\mathcal{U}$ random variable is helpful. We assume that all integrals and mathematical objects exist when referenced, as per usual.

Definition 2. Let g be a measurable function. A random variable $g(Y)$ will be said to be in the class of $\mathcal{U}$ random variables if and only if $E\{g(Y)\} = Av\{g(Y)\}$. Similarly, the same will be said w.r.t. $\mathbf{X} \in \mathbb{R}^k$ for $Y = g(X_1, \ldots, X_k)$ if and only if $EY = Av_{\mathbf{x}}(Y)$.

Definition 2 stipulates that a random variable is $\mathcal{U}$ class w.r.t. some space when its expected value is equal to its average functional value in that space. Stated in a probabilistic fashion, a variable Y is $\mathcal{U}$ class if one can take its expectation w.r.t. a uniform measure without changing its value. This type of random variable is ubiquitous in practice. For instance, any continuous uniform distribution or a normal distribution that has been symmetrically truncated about its expected value is a $\mathcal{U}$ random variable. Say $Y \sim Bern(\pi)$ s.t. $\pi = 2^{-1}$. Then Y is also in the $\mathcal{U}$ class, as is any $Y \sim Beta(\alpha, \beta)$ s.t. $\alpha = \beta$. Not all $\mathcal{U}$ random variables are symmetric, however. To illustrate this, construct a random variable Y with the following probabilities on $\mathcal{S}_Y = \{-5, 2, 3\} : \{14^{-1}5, 7^{-1}, 2^{-1}\}$. Then $EY = 14^{-1}(-25 + 4 + 21) = 0 = 3^{-1}(-5 + 2 + 3) = 0$. However, Y does not have a symmetric distribution.

A host of $\mathcal{U}$ properties has been investigated elsewhere [35]. To familiarize the reader, we provide a list of important ones in Table 1. Essentially, $\mathcal{U}$ variables are closely related in concept to sum-symmetry of the CDF and structured but uncorrelated deviations from uniformity. All bounded and symmetric random variables are in the $\mathcal{U}$ class, for instance, although symmetry is not a necessary condition, as previously demonstrated. Hence, all statistics that converge to a symmetric probability law behave more and more like $\mathcal{U}$ random variables as n becomes large. In an abuse of notation, we will say $Y \in \mathcal{U}$ if Y is in this class of random variable.

Out of these properties, we draw special attention to the concentration inequality: $\Pr(|S_n - ES_n| > \varepsilon) \leq 2\exp\{-\{\sum_{i=1}^{n} R_i^2\}^{-1}6\varepsilon^2\}$. The condition detailed in the table footnote is very mild and does not require independence. In fact, it can be true even when every single outcome variable in a sample is statistically dependent, insofar as the average correlation between those variables is mild, or μ_n is adequately bounded if this is not the case. Discussion on this assumption is also available elsewhere [35]. Put succinctly, a researcher can expect it to be fulfilled if each Y_i is symmetric – or at least relatively symmetric – and the joint distribution of the sample is biased away from n-tuples in the joint support that inflate S_n. We note that this is useful since these conditions often apply to the error distributions of statistics of interest, including those of linear regressions.

Table 1: Basic properties of $\mathcal{U}$ variables

Property	Description	Variable type	Conditions and definitions
$\sigma_{Y,f^{-1}(Y)} = 0$	No association with inverse density or mass function	A	$Y \sim f(y)$
$Y \in \mathcal{U} \Rightarrow cY \in \mathcal{U}, Y + c \in \mathcal{U}$	Algebraic closure	A	$c \in \mathbb{R}$
$\int_m^M F(y)\mathrm{d}y = \int_m^M S(y)\mathrm{d}y$	CDF is sum-symmetric	R, C, D	$S(y) = 1–F(y)$
$U = Y + \varepsilon$ s.t. $E(\varepsilon\|Y) = 0$	Uniform variables mean-preserving spreads of $\mathcal{U}$ variables	R, C	$U \sim \mathrm{Unif}(m, M)$, $f(y) \leq R^{-1}$ in left tail and unimodal
$\Pr(\|S_n–ES_n\| > \varepsilon) \leq 2\exp\{-\{\sum_{i=1}^n R_i^2\}^{-1}6\varepsilon^2\}$	Concentration inequality for $\mathcal{U}$ variables	R, C	$S_n = \sum_{i=1}^n Y_i$, $Y_i \in \mathcal{U}, C^*$
$\sum_{i=1}^M F(i) = EY$	Discrete CDF mean identity	R, D	$\mathcal{S}_Y = \{1, 2, \ldots, M\}$
$\sum_{i=1}^M F(i) - \sum_{i=1}^M S(i) = 1$	Discrete CDF sum-symmetry	R, D	$\mathcal{S}_Y = \{1, 2, \ldots, M\}$

A = All, R = Regular, C = Continuous, D = Discrete, $C^* = \max\{E(\exp\{sS_n\}), E(\exp\{–sS_n\})\} \leq Av_{\mathrm{y}}(\exp\{sS_n\})$, $s > 0$.

3.1 The Hoeffding bootstrap, continued

With $\mathcal{U}$ random variables introduced, we now define a slightly more conservative version of the Hoeffding bootstrap. This version is more justifiable for small samples, or when a larger body of statistical dependencies are expected to exist. It also presupposes that T_0 behaves more and more like a $\mathcal{U}$ class random variable as n grows larger: a proposition that is reasonable to assume for many statistics, especially random sums. In short, this strategy sets out to overpenalize to compensate for small sample bias, but then tries to taper this penalization with the efficiency gained by incorporating the $\mathcal{U}$ properties into the reasoning.

To this end, we essentially set $\phi = 2 \cdot \sqrt{6^{-1}\log(2/\alpha)}$. The increased penalization is present in the added factor of 2. However, the supposition of $\mathcal{U}$ status allows us to use the concentration inequality presented in Table 1 instead of Hoeffding's. All other things held equal, this strategy inflates the error around the estimate by an approximate factor of 1.15 in comparison to the first strategy when $\alpha = 0.05$. A longer justification for this construction is provided in the supplementary materials. We direct the reader there for more information.

Next, we introduce two simple propositions with direct practical or theoretical interest for causal inference. Proposition 3 follows directly from our main conditions and solves the problem of estimating $Av(Y_\delta|t)$ for variables in the $\mathcal{U}$ class since it allows for the replacement of the estimators of the previous section with the sample mean. Proposition 4 establishes an interesting sufficient condition for mean exchangeability. We only prove these statements for functional averages w.r.t. the range of Y. This is for conciseness. All results are easily extended to the excluded case.

Finally, recall that, although the notation is omitted, the following results also apply when the random variables are conditioned upon another vector of random variables $\mathbf{L}$, perhaps to facilitate the fulfillment of A4 or $\mathcal{U}$ status.

Proposition 3. *Suppose A1–A3 or A4. If* $Y_\delta|t \in \mathcal{U}$, *then* $E(Y_\delta|t) = Av\{Y(t)\}$.

Proof. The proof is again one line under the premises: $E(Y_\delta|t) = Av(Y_\delta|t) = Av\{Y(t)\}$. □

Again, the properties of plug-in estimators for equation (2.1) are not easy to discern in general and citations of the central limit theorem are also questionable or implausible. However, the properties of $\bar{Y}_\delta|t$ are exceptionally well-known. This largely solves the problem insofar as the sampling process secures a sample of $\mathcal{U}$ variables. Again, since Proposition 3 only requires the preservation of the support, this allows for an arbitrary distortion of the population distribution otherwise: a fact that is liberating w.r.t. study design and execution.

Note also that Proposition 3 is not as trivial as it seems. It is well known that the sample mean and mid-range estimate the same parameter when the underlying distribution is symmetric. However, it is false that all $\mathcal{U}$ random variables are symmetric. Hence, the $\mathcal{U}$ concept expands the universe where the sample mean can replace the mid-range. The next proposition establishes a new route to achieving mean exchangeability, as previously mentioned.

Proposition 4. *Suppose both $Y(t)$ and $Y_\delta|t$ are $\mathcal{U}$ random variables under A1–A3 or A4. Then $E\{Y(t)\} = E(Y_\delta|t)$.*

Proof. By our premises, the following string of equalities applies: $E\{Y(t)\} = Av\{Y(t)\} = Av(Y_\delta|t) = E(Y_\delta|t)$. □

Great care and energy of argument are often expended to establish that $E\{Y(t)\} = E(Y_\delta|t)$. Proposition 4 offers a new manner of doing so insofar as it is believed that the experimental distribution is sum-symmetric. Conditional or unconditional on some vector $\mathbf{L} = l$, insofar as the researcher is willing to posit that the experimental distribution is in the $\mathcal{U}$ class, all that is actually required is a sufficiently executed sampling process that preserves the support and induces *any* form of $\mathcal{U}$ status. Then it is implied that (conditional) mean exchangeability is achieved. Once more, since it seems plausible that $Y(t)$ or $Y(t)|\mathbf{l}$ can be mapped into a great number of $\mathcal{U}$ distributions on the same support via different sampling designs or conditioning, this result is potentially very useful.

For instance, if it is believed that the counterfactual distribution is symmetric, then a nonrejection of a statistical hypothesis of symmetry in the observed distribution can be supporting evidence that mean exchangeability is achieved. More generally, if the distance between the mid-range and the sample mean is small – and here one must be diligent in deciding what precisely defines the quality of this distance – this can also be construed as evidence. A researcher can also observe the behavior of the empirical CDF for visual confirmation. For regular random variables, the area below and above the curve should be approximately equal.

Before moving forward, however, more attention is owed to some possible implications of Proposition 4. First, we draw attention to the fact that C1–C3 are missing from the premises, and that it appears that A4 alone could suffice in conjunction with the added $\mathcal{U}$ condition. Since A4 implies at least a restricted form of positivity, this means – insofar as the counterfactual distribution is sum-symmetric – that other routes to the identification of expected causal effects exist in the absence of even the possibility of consistency or strong ignorability. Revisiting the test theory model helps to consider this matter further. Per this thought, observe the statement: $Y|t, \mathbf{L} = Y(t)|\mathbf{L} + \varepsilon$. Now, if there exists a single triplet $\{Y(t)|\mathbf{L}, Y|t, \mathbf{L}, \varepsilon\}$ s.t. $E\varepsilon = 0$ but $\Pr(\varepsilon = 0) \neq 1$, it subsequently follows that neither consistency nor strong ignorability are necessary conditions for mean exchangeability. One might expect $E\varepsilon = 0$ to be true with a nondegenerate ε distribution when unmeasured confounders impact the $T \to Y$ relationship nonlinearly.

For a more meaningful insight, however, we abandon enforcing mean exchangeability and only add a new constraint that $\varepsilon \perp\!\!\!\perp Y(t)|\mathbf{L}$. Stating this seems to signify that a researcher specified a $\mathbf{L}$ that was *almost* sufficient for achieving C1 and C2, but that independent noise, via measurement error or interference, persists to some degree. Recall that we have also accepted A4 as true. For simplicity, also say that $Y(t)|\mathbf{L}$ and ε are supported on compact intervals. From here, we can show that these suppositions jointly imply that $\varepsilon = 0$ almost surely. Since $\mathcal{S}_{Y(t)|\mathbf{L}} = [m, M]$ and $\mathcal{S}_\varepsilon = [-A, B]$, say, are both compact and $\varepsilon \perp\!\!\!\perp Y(t)|\mathbf{L}$, the support of the convolution of their densities is also compact. We also know by the Titchmarsh convolution theorem that this implies that $\mathcal{S}_{Y|t,\mathbf{L}} = [m - A, M + B]$. However, under A4, $\mathcal{S}_{Y|t,\mathbf{L}} = [m, M]$, which implies that $A = B = 0$ and that $\Pr(\varepsilon = 0) = 1$. Consequently, $Y|t, \mathbf{L} = Y(t)|\mathbf{L}$ and C1 and C2 are implied. Therefore, we can feel comfortable stating that A4 and a lack of omitted variable bias as defined are sufficient for deducing C1, C2, and at least a restricted form of positivity.

This thought experiment did not mention $\mathcal{U}$ status. However, we remark that the preservation of $\mathcal{U}$ status together with mean exchangeability might imply that omitted variable bias is absent, either as a standalone set of conditions or in conjunction with others that present themselves in relatively common circumstances. Such a state of affairs would supply C1 and C2 under the nice setup above. These facts have

yet to be determined. Additional investigation into the relationship between these conditions will doubtlessly be fruitful.

It also stands to mention that if either $Y|t, \mathbf{L}$ or $Y(t)|\mathbf{L}$ is a $\mathcal{U}$ random variable and we assume that A4 and mean exchangeability are valid conditions, then the other outcome is also a $\mathcal{U}$ random variable and the status is preserved by construction. To see this, say the shared support is $\mathcal{S} = [m, M]$ and note that A4 and mean exchangeability implies that $M - E(Y|t, \mathbf{L}) = M - E\{Y(t)|L\} = \int_S F\{y(t)|\mathbf{L}\}\mathrm{d}y = 2^{-1}R$. Hence, $\int_S F\{y|t, \mathbf{L}\}\mathrm{d}y = 2^{-1}R$, which implies that $\int_S F\{y|t, \mathbf{L}\}\mathrm{d}y = \int_S S\{y|t, \mathbf{L}\}\mathrm{d}y$. This achieves our result.

Finally, A4 is sufficient for establishing mean exchangeability when both the statistic being considered and its counterfactual partner converge in distribution to random variables with symmetric probability laws – or $\mathcal{U}$ probability laws more generally – as $n \to \infty$. For instance, observe some statistic $Z|t$. Since A4 applies, $Av(Z|t) = Av\{Z(t)\}$. However, since both $Z|t$ and $Z(t)$ converge to $\mathcal{U}$ random variables, this also implies that $E\{Z|t\} \approx E\{Z(t)\}$ for sufficiently large n. Once this situation is rendered more concrete by setting $Z|t = \bar{Y}|t$, it becomes apparent – in at least the context of additive statistics – that the presence of confounding, which negates mean exchangeability, also prevents the viability of A4 for the arithmetic means, or that it alters the dependency structure s.t. only one but not both statistics can converge in distribution to a $\mathcal{U}$ random variable. Again, more exploration is due. Since little scholarship has been dedicated to how confounding impacts the statistical dependencies between outcome variables, this is also an interesting route of investigation in its own right.

3.2 Counterfactual linear regression

Linear regression is a popular tool for causal and predictive inference. Inverse probability weighting (IPW) of the marginal structural model or standardization of the adjusted model are common methods for repurposing its framework for the former [6,14,52,53]. Doubly robust estimation, an example of which is the augmented inverse probability weighting estimator (AIPW), is also commonly employed for this purpose [54–56]. This section offers a brief review of these approaches before moving on to novel methods. The marginal structural model w.r.t. an event $\{T = t\}$ is defined as follows:

$$E\{Y(t)\} = \kappa_0 + \beta_1 t. \tag{3.1}$$

The adjusted model is defined similarly, but in conjunction with a vector of adjusting variables $\mathbf{L}$ under the auspices that C1–C3 are true:

$$E(Y|t, \mathbf{L}) = E\{Y(t)|\mathbf{L}\} = \beta_{0_*} + \beta_1 t + \mathbf{L}\boldsymbol{\beta}. \tag{3.2}$$

Standardization then yields that $E\{E\{Y(t)|\mathbf{L}\}\} = \beta_{0_*} + \beta_1 t + E\{\mathbf{L}\}\boldsymbol{\beta} = E\{Y(t)\}$. Under this model, the marginal expected causal effect is: $E\{Y(1)\} - E\{Y(0)\} = \beta_1$. This effect is typically estimated using empirical plugins from the fitted multiple linear regression model and standardization [6,57]. As aforementioned, the identification of a vector $\mathbf{L}$ that achieves C1–C3 is nontrivial, as is the felicitous specification of equation (3.2).

These difficulties transport to IPW models. Transitioning to this topic, say $e(\mathbf{L}) = \Pr(T = 1|\mathbf{L})$ is a propensity score model that captures the probabilistic mechanisms underlying treatment exposure. Provided this, it is subsequently true that $E\{e^{-1}(\mathbf{L})(Y \cdot T) - \{1 - e(\mathbf{L})\}^{-1}\{Y \cdot (1 - T)\}\} = E\{Y(1)\} - E\{Y(0)\}$, again under C1-C3. All things merry and well, an IPW estimator such as $\hat{\Delta}_{1,0;\mathrm{IPW}} = n^{-1}\sum_{i\in I}\{\hat{e}^{-1}(\mathbf{L}_i)(Y_i \cdot T_i) - \{1 - \hat{e}(\mathbf{L}_i)\}^{-1} \cdot Y_i \cdot (1 - T_i)\}$ is statistically consistent for the targeted estimand, where $\hat{e}(\mathbf{L})$ is usually estimated via logistic regression or some other semi-parametric method. The problem is, very rarely, are all things merry and well. As aforementioned, like the specification of equation (3.2), approximating $e(\mathbf{L})$ is an arduous task that is burdened by doubt. This state of affairs is again exacerbated by the strictness of conditions C1–C3, and in particular C2.

Doubly robust estimation methods attempt to address the issue of incorrect model specification. In essence, they augment existing estimators for causal effects with the features of another s.t. the resulting estimator is statistically consistent for the identified estimand as long as *at least one* of the underlying models is correctly specified. Say $\hat{Y}_i(t)$ is an estimator of a mean outcome from equation (3.2) that follows

the plug-in parametric g-formula. Then one example of a doubly robust statistic is the AIPW estimator: $\hat{\Delta}_{1,0;\text{AIPW}} = n^{-1}\sum_{i\in I}\{\hat{Y}_i(1) + \hat{e}^{-1}(\mathbf{L}_i)T_i\{Y_i - \hat{Y}_i(1)\}\} - n^{-1}\sum_{i\in I}\{\hat{Y}_i(0) + \{1 - \hat{e}(\mathbf{L}_i)\}^{-1}(1 - T_i)\{Y_i - \hat{Y}_i(0)\}\}$.

Pertinently, all of these models face another shared difficulty: informative sampling. This is because they all revolve around expected values. Unfortunately, expected values are susceptible to alterations in the distribution (s) of the outcome variables, which result from the particularities of a study design and its sampling mechanism. For instance, for equation (3.2), what is actually estimated is: $E_\delta(Y|t, \mathbf{L}) = \alpha_0 + \alpha_1 t + \mathbf{L}_\delta\boldsymbol{\alpha}$. Hence, even when C1–C3 are met, it is $E_\delta\{Y(t)\}$ that is identifiable in the absence of noninformative sampling or additional constraints. $E_\delta\{Y(t)\}$, however, might not be the target of interest.

We now use concepts from the previous sections to demonstrate that the core assumptions of linear regression for predictive (associative) inference are sufficient for the identification of causal parameters in conjunction with A4 and a causal theory. In essence, we prove that functional average causal estimands remain identified and estimable using elementary methods alone, and even when C1–C3 do not hold and informative sampling is present. To this end, we specify the data-generating mechanism for the linear model in equation (3.3) with $\mathbf{L} = \mathbf{l}$ fixed.

$$Y_{\delta_i} = \alpha_0 + \alpha_1 t_i + \mathbf{l}_i\boldsymbol{\alpha} + \varepsilon_{\delta_i}. \tag{3.3}$$

Traditionally, equation (3.3) requires that $E_\delta(\varepsilon_i|t_i, \mathbf{l}_i) = 0$ for $\forall i$ when predictive inference is the goal and all covariate patterns of interest are fixed. This is weaker than strict exogeneity, which requires that $E_\delta(\varepsilon_i|T_i, \mathbf{L}_i) = 0$ when T_i and $\mathbf{L}_i$ are stochastic. Using standardization requires a slightly weaker form of strict exogeneity that is conditioned on all treatment contrasts of interest since the method averages over $\mathbf{L}_\delta$: $E_\delta(\varepsilon_i|t_i, \mathbf{L}_i) = 0$ for $\forall i$. Otherwise, for finite sample inference, the second core assumption is that $\varepsilon_{\delta_i} \sim N(0, \sigma_i^2)$ for $\forall i$.

The first core assumption is commonly evaluated using the predicted versus residual plot. Under the working proposition of valid specification, this plot should demonstrate an approximately symmetric scatter of the residuals about the horizontal zero line for any arbitrarily small neighborhood around any predicted point on the x-axis.

To make use of these traditional conditions, we must first make an inconsequential adjustment to the assumption of normality. We are working within a universe of bounded random variables. Consequently, the ε_{δ_i} of equation (3.3) cannot be normally distributed. This is no great loss for four related reasons. First, in a grand majority of scientific investigations, Y_i is bounded. For example, if each Y_i is a measurement of a person's blood pressure, it is impossible for it to be less than zero or greater than an arbitrary real number. Its distribution cannot be supported on a set that is equal to $\mathbb{R}$. This automatically implies that the ε_i cannot truly be normally distributed. In these situations, when statisticians assume normality, it is intended as a feasible approximation that results in negligible error, and that is fecund mathematically.

The second reason is similar to the first. Even if someone wishes to insist that Y_i is supported on the entire real line, Y_{δ_i} often cannot be due to the intrinsic limitations of measurement and observation. Third, as hinted in the first reason, the assumption of normality can be replaced with the assumption that ε_{δ_i} has a CDF $F_i(e_\delta)$ of a normal distribution that has been symmetrically truncated around zero. This is equivalent to positing that each ε_{δ_i} is related some variable $Z_i \sim N(0, \sigma_{Z_i}^2)$ s.t. for an (almost) arbitrarily small $\tau > 0$, $\Pr(-M_i \le Z_i \le M_i) = 1 - \tau$ and $F_i(e_\delta) = (1 - \tau)^{-1}\Phi_{Z_i}(e_\delta)$ on $[-M_i, M_i]$. Provided this setup, the bias that results from treating ε_{δ_i} as strictly normally distributed for mathematical convenience is unimportant, especially since one does not need to identify $\sigma_{Z_i}^2$.

The fourth and last reason, which motivates the next proposition, is related to the requirement of symmetric scatter in the residual versus fitted plot. A symmetrically truncated normal distribution is a special case of a $\mathcal{U}$ random variable. Moreover, when a continuous random variable has an expected value of zero, all that is required for regular $\mathcal{U}$ status is for it to be supported on a symmetric interval $[-M_i, M_i]$.

Hence, the typical set of assumptions already employed for fixed linear regression already requires that each $\varepsilon_{\delta_i} \in \mathcal{U}$. In addition, we note that positing that $\varepsilon_{\delta_i} \in \mathcal{U}$ $\forall i$ is a fundamentally weaker assumption than (symmetrically truncated) normality. Under this milder condition, a researcher only needs to verify that the residual versus predicted plot is (approximately) symmetrically *supported* around zero about any neighborhood of predicted values. The behavior of the scatter within any neighborhood is otherwise unimportant,

insofar as it reasonably justifies that the expected value is also zero. Nevertheless, it is apropos to state that, if only this milder condition is supposed, then the concentration inequality of Table 1 or a central limit theorem are required for the construction of confidence intervals. Of course, under copious amounts of dependencies, a central limit will not necessarily apply.

We now present a useful main result in Proposition 5, although it is technically a more detailed case of Proposition 3. The extra assumption of regularity is not strictly necessary.

Proposition 5. *Assume A4. Say* $Y_\delta|t,\mathbf{l} = g(t,\mathbf{l},\boldsymbol{\alpha}) + \varepsilon_\delta$ *for some measurable (possibly monotonic) function* g*. Suppose each* ε_δ *is regular,* $E_\delta(\varepsilon|t,\mathbf{l}) = 0$*, and* $\varepsilon_\delta \in \mathcal{U}$ *for* $\forall t,\mathbf{l}$ *fixed. Then* $E_\delta(Y|t,\mathbf{l}) = Av\{Y(t)|\mathbf{l}\}$.

Proof. Let $t,\mathbf{l}$ be arbitrary. Then $E_\delta(Y|t,\mathbf{l}) = g(t,\mathbf{l},\boldsymbol{\beta})$ since $E_\delta(\varepsilon|t,\mathbf{l}) = 0$.

For an arbitrary bounded random variable Z, say $\min(Z) = \min_{z\in\mathcal{S}_Z}(z)$ and $\max(Z) = \max_{z\in\mathcal{S}_Z}(z)$, the greatest lower and least upper bounds of the closed set $\mathcal{S}_Z$, respectively. Since $g(t,\mathbf{l},\boldsymbol{\beta})$ is a constant, it follows that $\min(Y_\delta|t,\mathbf{l}) = g(t,\mathbf{l},\boldsymbol{\beta}) + \min(\varepsilon_\delta)$ and $\max(Y_\delta|t,\mathbf{l}) = g(t,\mathbf{l},\boldsymbol{\beta}) + \max(\varepsilon_\delta)$. Moreover, since each ε_δ is regular, then each $Y_\delta|t,\mathbf{l}$ is also obviously regular. From here:

$$\begin{gathered}\min(Y_\delta|t,\mathbf{l}) + \max(Y_\delta|t,\mathbf{l}) = 2g(t,\mathbf{l},\boldsymbol{\beta}) + \min(\varepsilon_\delta) + \max(\varepsilon_\delta) \Rightarrow \\ 2^{-1}\{\min(Y_\delta|t,\mathbf{l}) + \max(Y_\delta|t,\mathbf{l})\} = g(t,\mathbf{l},\boldsymbol{\beta}) + 2^{-1}\{\min(\varepsilon_\delta) + \max(\varepsilon_\delta)\} \Rightarrow \\ Av(Y_\delta|t,\mathbf{l}) = g(t,\mathbf{l},\boldsymbol{\beta}) + 0 \Rightarrow \\ E_\delta(Y|t,\mathbf{l}) = Av(Y_\delta|t,\mathbf{l}) = Av\{Y(t)|\mathbf{l}\}\end{gathered}$$

The third line follows from regularity and the fact that $\varepsilon_\delta \in \mathcal{U}$ for $\forall t,\mathbf{l}$ fixed. The last line follows from substitution and A4. □

Set $g\{t,\mathbf{l},(\alpha_0,\alpha_1,\boldsymbol{\alpha})\} = \alpha_0 + \alpha_1 t + \mathbf{l}\boldsymbol{\alpha}$ to recover equation (3.3). Under the assumption that $(t',\mathbf{l})$ and $(t'',\mathbf{l})$ have both been fixed, where the values t' and t'' represent the treatment values to be contrasted, Proposition 5 implies that $\alpha_1 \propto Av\{Y(t')|\mathbf{l}\} - Av\{Y(t'')|\mathbf{l}\}$: the difference in the *average* values of $Y(T)|\mathbf{l}$ when $T = t'$ and $T = t''$. When $(t',\mathbf{l})$ and $(t'',\mathbf{l})$ have not been fixed but at least all treatment values have been, i.e., when the researcher did not fix $\mathbf{l}$ for all contrasts of scientific interest, the previous statement still holds in general when $E_\delta(\varepsilon|t,\mathbf{L}) = 0$ or $\forall t$ involved. The proof of Proposition 5 would only need to use this stronger statement with no further change. If the researcher wishes to reason about contrasts of T that have not been fixed as well, then the (nontrivial) assumption that $E_\delta(\varepsilon|T,\mathbf{L}) = 0$ suffices.

This proves that conditions that are weaker than those traditionally supposed for making inferences about associations are sufficient for inference w.r.t. causal parameters. A statistician can target these parameters using linear regression under (almost) arbitrary sampling bias, insofar as linearity and $\mathcal{U}$ status in the error distributions are feasibly defensible w.r.t. the sample measures, at least the supports have been preserved, and she can defend the causal assertions. Importantly, A4 can be weakened further since we truly only require the preservation of functional averages.

We choose to highlight five additional points of interest. The first one is about the interpretation of α_1. It uses similar language to current interpretations of regression coefficients. However, care is due. Although the word "average" is often used for interpreting the coefficients of typical linear regressions, this is imprecise slang for the change in expected value. In distinction, the word "average" is precise for the functional average since it is a uniform averaging over all possible values that the outcome can materialize.

The second point is a caveat. Proposition 5 enables reasoning about counterfactual conditional functional averages in high-dimensional settings. Often, however, the researcher cares mostly about a marginal estimate, and – although they can serve a similar purpose – $E\{Av\{Y(t)|\mathbf{L}\}\} \neq Av\{Y(t)\}$ and $Av\{Av\{Y(t)|\mathbf{L}\}\} \neq Av\{Y(t)\}$ generally. For clarity, we designate L as a single discrete covariate WLOG to exemplify these estimands below:

$$E\{Av\{Y(t)|L\}\} = \sum_{l\in\mathcal{S}_L}\{R_{Y|l}^{-1}\int_{\mathcal{S}_{Y|l}} y\mathrm{d}y\}f(l),$$

$$Av\{Av\{Y(t)|L\}\} = |\mathcal{S}_L|^{-1}\sum_{l\in\mathcal{S}_L}\{R_{Y|l}^{-1}\int_{\mathcal{S}_{Y|l}} y\mathrm{d}y\}.$$

Although they might not reduce perfectly to an isolated marginal effect, this does not signify that these parameters do not possess scientific meaning or that they are not attractive in any way – and this is the core of our third point. For example, $E\{Av\{Y(t')|L\}\} - E\{Av\{Y(t'')|L\}\}$ can still be interpreted as an *expected* functional average effect over $\mathbf{L}$. Furthermore, unlike iterated expectations, $Av\{Av\{Y(t)|L\}\}$ does not strictly require full positivity since it is always a well-defined average over those $L = l$ that are observable when $T \in \{t', t''\}$, the set of contrasted treatment exposures. This is also why we observed that A4 at least implies restricted positivity, since it requires that $f(t|\mathbf{l}) > 0$ for some variable $\mathbf{L}_*$ that possibly represents a censored version of $\mathbf{L}$. Once again, this censuring does not matter insofar as averaging over the $\mathbf{L} = \mathbf{l}$ of scientific interest is possible for the targeted treatment exposures. Nevertheless, one special circumstance when the identity $E\{Av\{Y(t)|\mathbf{L}\}\} = Av\{Y(t)\}$ does hold is when the premises of Proposition 4 hold: a point we return to shortly.

Expected or uniform averaging of functional averages over the strata of $\mathbf{L}$ does possess a succinct meaning when the underlying model is linear. To see this, again appreciate the model, assuming it is validly specified in the spirit of Proposition 5: $Av\{Y(t)|\mathbf{L}\} = \alpha_0 + \alpha_1 t + \mathbf{L}\boldsymbol{\beta}$. Obviously then, if this is the case, then $Av\{Y(t')|\mathbf{L}\} - Av\{Y(t'')|\mathbf{L}\} = \alpha_1(t' - t'')$. Since this is constant, $E\{Av\{Y(t')|\mathbf{L}\} - Av\{Y(t'')|\mathbf{L}\}\} = Av\{Av\{Y(t')|\mathbf{L}\} - Av\{Y(t'')|\mathbf{L}\}\} = \alpha_1(t' - t'')$. In short, all averaged estimands are proportional to the regression coefficient.

Point four requires us to revisit the AIPW estimator. Although this estimator has been used to target a different estimand historically, it is a statistically consistent estimator of $E\{Av\{Y(1)|\mathbf{L}\} - Av\{Y(0)|\mathbf{L}\}\}$ under mild dependency conditions if the outcome model is validly specified, C1–C3 hold, and the premises of Propositions 4 or 5 do as well. This is because – under these premises – $E\{Av\{Y(1)|\mathbf{L}\} - Av\{Y(0)|\mathbf{L}\}\} = E\{E\{Y(1)|\mathbf{L}\} - E\{Y(0)|\mathbf{L}\}\} = E\{Y(1)\} - E\{Y(0)\}$. Otherwise, the AIPW estimator and functional average estimators target different estimands.

Table 2: Estimators and ladders of assumptions

Estimator	Assumption ladder	Estimand
$\hat{\mathrm{MR}}|t, l$	(1): $Y|t, l = Y(t)|l + \tau$, $\mathcal{S}_{Y(t)|l+\tau} = \mathcal{S}_{Y(t)|l} \times \mathcal{S}_\tau$, $Av(\tau) = 0$	$Av\{Y(t)|l\}$
	(2): A1–A3 or A4	
$\lvert\mathcal{S}_{L_\delta}\rvert^{-1}\sum_{l\in\mathcal{S}_{L_\delta}}\hat{\mathrm{MR}}|t, l$	(1) or (2)	$Av\{Av\{Y(t)|L_\delta\}\}$
$\sum_{l\in\mathcal{S}_{L_\delta}}\hat{\mathrm{MR}}|t, l \cdot \hat{f}(L_\delta = l)$	(3): (1) or (2), $E_\delta(\cdot) = E(\cdot)$, C3	$E\{Av\{Y(t)|L\}\}$
	(3) and $Y(t)|l \in \mathcal{U}$ for $\forall l \in \mathcal{S}_{L_\delta}$	$E\{Y(t)\}$
$\bar{Y}|t, l$	(1): $\bar{Y}|t, l = \bar{Y}(t)|l + \tau$, $\mathcal{S}_{\bar{Y}(t)|l+\tau} = \mathcal{S}_{\bar{Y}(t)|l} \times \mathcal{S}_\tau$, $Av(\tau) = 0$	$Av\{\bar{Y}(t)|l\}$
	(2): A1–A3, A4 for $\bar{Y}|t, l$ and $\bar{Y}|t, l \in \mathcal{U}$ for large n	
	(3) A1–A3 or A4 for $Y|t, l$ and $Y|t, l \in \mathcal{U}$	$Av\{Y(t)|l\}$
	(3) and $Y(t)|l \in \mathcal{U}$	$E\{Y(t)|l\}$
$\lvert\mathcal{S}_{L_\delta}\rvert^{-1}\sum_{l\in\mathcal{S}_{L_\delta}}\bar{Y}|t, l$	(1) or (2)	$Av\{Av\{\bar{Y}(t)|L_\delta\}\}$
	(4): (1) or (2) and $\bar{Y}(t)|l \in \mathcal{U}$ for $\forall l \in \mathcal{S}_{L_\delta}$	$Av\{E\{\bar{Y}(t)|L_\delta\}\}$
	(4) and $E\{\bar{Y}(t)|L_\delta\} \in \mathcal{U}$	$E\{\bar{Y}(t)\}$
$\sum_{l\in\mathcal{S}_{L_\delta}}\bar{Y}|t, l \cdot \hat{f}(L_\delta = l)$	(5): (1) or (2), $E_\delta(\cdot) = E(\cdot)$, C3	$E\{Av\{\bar{Y}(t)|L\}\}$
	(5) and $\bar{Y}(t)|l \in \mathcal{U}$ for $\forall l \in \mathcal{S}_L$	$E\{\bar{Y}(t)\}$
$\hat{Y}|t, \mathbf{l}$	(1): $Y|t, \mathbf{l} = Y(t)|\mathbf{l} + \tau$, $\mathcal{S}_{Y(t)|\mathbf{l}+\tau} = \mathcal{S}_{Y(t)|\mathbf{l}} \times \mathcal{S}_\tau$,	$Av\{Y(t)|\mathbf{l}\}$
	$Av(\tau) = 0$, $Y|t, \mathbf{l} = g(t, \mathbf{l}, \boldsymbol{\alpha}) + \varepsilon$, $E(\varepsilon|t, \mathbf{l}) = 0$, $\varepsilon \in \mathcal{U}$	
	(2): A1–A3 or A4, $Y|t, \mathbf{l} = g(t, \mathbf{l}, \boldsymbol{\alpha}) + \varepsilon$, $E(\varepsilon|t, \mathbf{l}) = 0$, $\varepsilon \in \mathcal{U}$	
	(1) or (2) and $Y(t)|\mathbf{l} \in \mathcal{U}$	$E\{Y(t)|\mathbf{l}\}$
$\lvert\mathcal{S}_{\mathbf{L}_\delta}\rvert^{-1}\sum_{\mathbf{l}\in\mathcal{S}_{\mathbf{L}_\delta}}\hat{Y}|t, \mathbf{l}$	(1) or (2)	$Av\{Av\{Y(t)|\mathbf{L}_\delta\}\}$
	(1) or (2), $\mathcal{S}_{\mathbf{L}_\delta} = \mathcal{S}_{\mathbf{L}}$, $Y(t)|\mathbf{l} \in \mathcal{U}$ for $\forall\ \mathbf{l} \in \mathcal{S}_{\mathbf{L}}$, $E\{Y(t)|\mathbf{L}_\delta\} \in \mathcal{U}$	$E\{Y(t)\}$
$\sum_{\mathbf{l}\in\mathcal{S}_{\mathbf{L}_\delta}}\hat{Y}|t, \mathbf{l} \cdot \hat{f}(\mathbf{L}_\delta = \mathbf{l})$	(3): (1) or (2), $E_\delta(\cdot) = E(\cdot)$, C3	$E\{Av\{Y(t)|\mathbf{L}\}\}$
	(3) and $Y(t)|\mathbf{l} \in \mathcal{U}$ for $\forall \mathbf{l} \in \mathcal{S}_{\mathbf{L}}$	$E\{Y(t)\}$

That is not to assert, however, that the AIPW estimator is not useful for comparison or the testing of hypotheses. Recall: standard linear regressions already assume $\mathcal{U}$ errors. Since the AIPW estimator partially relies on a well-specified outcome model, the $\mathcal{U}$ assumption is often intrinsic to its use. Using the AIPW estimator, however, requires C1–C3, which guarantees $A4$. Therefore, in most contexts where it is employed, it is implicitly assumed that $Y(t)|\mathbf{L} \in \mathcal{U}$ and $Y|t, \mathbf{L} \in \mathcal{U}$. In conclusion, the AIPW estimator and $E\{Av\{Y(1)|\mathbf{L}\} - Av\{Y(0)|\mathbf{L}\}\}$ do target the same estimand in typical practice settings. This is useful to know. Researchers can then commence their analysis under A1–A3 or A4 to identify one level of causal estimands, including $E\{Av\{Y(1)|\mathbf{L}\} - Av\{Y(0)|\mathbf{L}\}\}$. If the distance between the latter and the AIPW estimator is sufficiently small, then this is information that can at least be used to support the notion that C1–C3 has been achieved. This same point stands for the plug-in g-formula estimator. In this sense, one can justify stepping upwards on the ladder of assumptions to more defensibly assert the identification of traditional causal estimands.

Table 2 offers a summary of some of these ladders of assumptions w.r.t. a subset of estimators explored in this article. It does not exhaust all possible assumptions and estimands in the context of functional averages and their close relationship to expected values. Nevertheless, it covers much ground and requires a short exposition for clarity. Each subsection of rows offers information on a particular family of estimators and commences with two sets of assumptions labeled (1) and (2). The first set abbreviates a version of the "test theory" basis for functional average preservation while the second starts from the main assumptions of this article. From there, additional assumptions are added to either grouping. These assumptions allow for the transformation of the targeted estimand into one that is closer to the traditional one. In general, assumption groupings that rest on the ladder's lower rungs are easier to achieve, although this might not be true in particular research contexts. Also, the rows of Table 2 that correspond to probabilistic averaging over strata include C3 as an added assumption. Technically, this is unnecessary. However, C3 has been included since positivity is required for differences in counterfactual expected values to be well-defined. Recall also that $E_\delta(\cdot)$ denotes an expectation that is taken w.r.t. the sample distribution. When the table asserts that $E_\delta(\cdot) = E(\cdot)$ is required, it is stating that the expected value w.r.t. the sample distribution must be unbiased for the population expected value. This is a strictly weaker condition than noninformative sampling, as previously discussed. Furthermore, although all sample estimators technically require δ notation, we have only included it in a few places. This is for ease of reading and to draw attention to this fact in more critical places, such as where expected values are involved or a uniform averaging exists over the support of a conditioning variable.

The last point is concise to state. The analysis of variance (ANOVA) is a special case of the linear regression model presented. It is an elementary tool that is ubiquitous in research. All prior discussion and Proposition 5 therefore apply to ANOVA procedures under conditions that are already stipulated. Hence, insofar as A1–A3 or A4 are defensible, this means that a plethora of prior work can be re-interpreted with a restricted causal lens in partnership with a much more general structural causal model that allows for unmeasured confounders.

4 Monte Carlo simulations

Before we apply our strategy to real data, we illustrate its utility with a set of Monte Carlo simulations that show how functional average and $\mathcal{U}$ concepts are useful for causal inference. For simplicity, we proceed with noninformative sampling conditions that presuppose mutual independence. All simulations use $M = 1{,}000$ iterations for sample sizes $n \in \{500,\ 2{,}500,\ 5{,}000,\ 10{,}000\}$. Furthermore, all constructed $1 - \alpha$ confidence sets use $\alpha = 0.05$. Three main simulation experiments are provided in total. The first examines the behavior of basic functional average estimators for symmetric and nonsymmetric distributions. The second and third simulations demonstrate that causal effects can be consistently estimated without controlling for confounding. All experiments examine the performance of Hoeffding style bootstrapping procedures.

4.1 Univariate functional average estimation

The first simulation of this experiment examines the performance of $\hat{MR}$ for three truncated normal distributions: $Y_1 \sim TN(m = 0, M = 20, \mu = 10, \sigma = 5)$, $Y_2 \sim TN(m = 0, M = 15, \mu = 10, \sigma = 3)$, and $Y_3 \sim TN(m = 0, M = 15, \mu = 5, \sigma = 3)$.

The first random variable Y_1 is in the $\mathcal{U}$ class and hence $\theta_1 = Av(Y_1) = EY_1 = 10$. However, Y_1 and Y_2 are not $\mathcal{U}$ variables. Their distributions are skewed with tails that impact convergence behavior. For these variables, $\theta = Av(Y) = 7.5$.

Hoeffding bootstrap style confidence sets ($\hat{CI}_H$) are constructed as described in Section 2. To contrast its performance, we also use an m-out-of-n bootstrap. Again, an important requirement of the m-out-of-n bootstrap is that $m \to \infty$, but $n^{-1}m \to 0$. To meet this criteria, we set $m = r(\sqrt{n})$ since this setting produces relatively conservative results. Hence, if it fails to perform well, this highlights the utility of the Hoeffding procedure. Percentile confidence sets ($\hat{CI}_{p,m}$) are constructed from the m-out-of-n bootstraps. For reference we also construct Hoeffding style confidence sets ($\hat{CI}_{H,m}$) from this process. Importantly, all bootstrap procedures make use of simple random sampling with replacement and only 500 bootstrap samples. Although suboptimal, a low number of bootstrap samples is used to limit computational burden. A decent performance of the Hoeffding bootstrap at $B = 500$ is still a good indicator. Empirical coverage rates are estimated with $EC = M^{-1}\sum_{i=1}^{M} 1_{\theta \in \hat{CI}_i}$ WLOG.

Table 3 presents the results of this experiment. Importantly, all values in the tables henceforth represent the arithmetic average of simulated objects, including the endpoints of confidence sets.

These results demonstrate that $\hat{MR}$ behaves as intended. It is unbiased for the symmetric distribution and convergence behavior – albeit slow – is observable w.r.t. the target parameters for the asymmetric distributions that exhibit more problematic tail behavior. Importantly, the Hoeffding-style bootstrap also appears to behave as intended. Although it failed to uphold nominal coverage values for the skewed distributions at $n = 500$, it quickly overcame this behavior to provide conservative empirical coverage for θ. Also, the difficulties of efficiently estimating $Av(Y)$ are evident for non-$\mathcal{U}$ variables, highlighting the utility of this type of variable. Notably, the m-out-of-n percentile interval did not perform well even when m was $O(\sqrt{n})$.

The second experiment is for discrete variables. Like before, we use three different truncated normal distributions: $Y_1 \sim TN(m = 0, M = 40, \mu = 20, \sigma = 5)$, $Y_2 \sim TN(m = 0, M = 40, \mu = 25, \sigma = 8)$, and $Y_3 \sim TN(m = 0, M = 40, \mu = 15, \sigma = 8)$. These random variables are uniformly rounded to the nearest integer to induce discreteness. Here, we contrast $\hat{MR}$ with $\hat{Av}$, the discrete plug-in for equation (2.1). Hoeffding style bootstraps are again employed to construct confidence sets. However, now we use a $\mathcal{U}$ approximation in accordance with Section 3. Even if $\mathcal{U}$ status does not hold exactly, insofar as convergence behavior to a constant holds, the method should remain robust.

Results for this simulation are available in Table 4. We no longer examine the performance of the m-out of-n bootstrap.

Table 3: Functional average estimation, continuous

$Av(Y)$	n	$\hat{MR}$	$\hat{CI}_{H,m}$	$EC_{H,m}$	$\hat{CI}_{p,m}$	$EC_{p,m}$	$\hat{CI}_H$	EC_H
$\theta = 10$	500	10.01	(1.85, 18.15)	1	(8.02, 11.98)	1	(8.87, 11.13)	1
	2,500	10	(4.5, 15.5)	—	(8.72, 11.27)	—	(9.72, 10.28)	—
	5,000	10	(5.42, 14.58)	—	(8.97, 11.03)	—	(9.86, 10.14)	—
	10,000	10	(6.26, 13.74)	—	(9.17, 10.83)	—	(9.93, 10.07)	—
$\theta = 7.5$	500	8.11	(2.1, 14.12)	1	(7.58, 10.64)	0.40	(6.51, 9.7)	0.88
	2,500	7.74	(2.98, 12.5)	—	(7.6, 10.09)	0.30	(6.82, 8.65)	0.94
	5,000	7.63	(3.31, 11.96)	—	(7.59, 9.87)	0.25	(6.96, 8.31)	0.97
	10,000	7.58	(3.7, 11.46)	—	(7.58, 9.67)	0.21	(7.13, 8.03)	0.98
$\theta = 7.5$	500	6.91	(0.88, 12.94)	1	(4.37, 7.44)	0.43	(5.31, 8.51)	0.90
	2500	7.26	(2.46, 12.06)	—	(4.91, 7.4)	0.30	(6.34, 8.18)	0.95
	5,000	7.37	(3.04, 11.69)	—	(5.13, 7.41)	0.23	(6.69, 8.04)	0.97
	10,000	7.42	(3.52, 11.33)	—	(5.33, 7.42)	0.18	(6.97, 7.87)	0.97

† Each θ corresponds to the functional average of the TN distributions introduced above; " – " indicates that the value is the same as the first row.

The results of the discrete simulation largely match those of the preceding continuous one. The plug-in estimator performed more favorably than the sample mid-range only for $r(Y_1)$: the distribution with the lightest tails. Importantly, although many simulations demonstrated departures from $\mathcal{U}$ status, the coverage remained robust due to the conservative nature of the Hoeffding bootstrap. Further details pertaining to this fact are available in the supplementary material.

4.2 Causal inference with functional averages

Next, we demonstrate a classical case of confounding where the functional average treatment effect is identifiable and statistically consistently estimable without any adjustment. We use the following variables for this simulation: $C \sim Bern(0.5)$, $T \sim Bern(0.3 + 0.5C)$, and $e \sim TN(m = -50, M = 50, \mu = 0, \sigma = \tau)$. Moreover, we will say that $Y(1) = 110 + 50C + e$, $Y(0) = 100 + 50C + e$, and $Y = Y(1)T + (1 - T)Y(0)$.

The confounding variable is C, which is present in half of the theoretical population on average. When $C = 1$, the probability of allocation to treatment is larger. The structure of the confounding also preserves the support of the counterfactual distribution w.r.t. the observed one. We also set τ to 5 then 25 to demonstrate the performance of a functional average estimator when the tail probabilities are thin or heavy. Notably, this experiment is also structured s.t. A1–A3 and therefore A4 all hold.

We contrast simple linear regression with $\hat{MR}$ for estimating $\Delta = Av\{Y(1)\} - Av\{Y(0)\} = 10$. Here, the mid-range estimator of Δ is $\hat{\Delta}_{\text{MR}} = 2^{-1}\{Y_{(1)}|1 + Y_{(n_1)}|1\} - 2^{-1}\{Y_{(1)}|0 + Y_{(n_0)}|0\}$. This simulation is executed WLOG since it can be implicitly assumed that the researcher has stratified on some set of variables – such as propensity scores – to limit confounding bias or achieve the equality of supports. Confidence sets and estimators of the empirical coverage are all otherwise constructed as previously explored using the Hoeffing bootstraps of Section 2. Power is estimated by $\text{EP}_H = M^{-1}\sum_{i=1}^{M} 1_{0 \notin \hat{CI}_{H_i}}$. Table 5 has the results.

As expected, $\hat{\Delta}_{\text{OLS}}$ is a confounded estimator of $E\{Y(1)\} - E\{Y(0)\} = \Delta_{1,0} = 10$. However, this is not the case for $\hat{\Delta}_{\text{MR}}$, which demonstrates convergence behavior towards the true parameter value, albeit at a suboptimal rate. Reiteration of a poignant fact here is valuable: $\hat{MR}$ demonstrates this behavior without adjustment. Moreover, although the sampling conditions supposed here were noninformative, this was unnecessary. Insofar as the supports are preserved, informative sampling conditions are unimportant w.r.t. statistical consistency, although they might impact convergence rates. These results are also consistent with prior discussions on the mid-range. We expect the mid-range to possess more favorable properties when nonnegligible mass or density rests in the tails, which is the case when $\tau = 25$ for these simulations.

Next, we present a restricted discrete analogue to the last experiment. C and T retain their definitions, but $\varepsilon \sim Binom(\tau, 2^{-1})$ for $\tau \in \{30, 50\}$. Otherwise, $Y(0) = 10C + e$, $Y(1) = Y(0) + 5$, and $Y = Y(1)T + (1 - T)Y(0)$

Table 4: Functional average estimation, discrete

$Av(Y) = 20$	n	$\hat{Av}$	$\hat{CI}_H$	EC_H	$\hat{MR}$	$\hat{CI}_H$	EC_H
$r(Y_1)$	500	20.012	(15.21, 24.82)	1	20.044	(13.88, 26.21)	1
	2500	19.992	(15.84, 24.14)	—	19.969	(14.93, 25.01)	—
	5,000	20.014	(16.18, 23.85)	—	20.023	(15.58, 24.47)	—
	10,000	20.020	(16.39, 23.65)	—	20.016	(16, 24.04)	
$r(Y_2)$	500	21.921	(17.62, 26.22)	0.971	21.218	(16.48, 25.96)	0.955
	2,500	20.490	(18.19, 22.79)	0.980	20.365	(18.17, 22.56)	0.963
	5,000	20.194	(18.55, 21.83)	0.985	20.169	(18.67, 21.67)	0.975
	10,000	20.051	(19, 21.1)	0.972	20.050	(19.06, 21.04)	0.957
$r(Y_3)$	500	18.053	(13.8, 22.31)	0.963	18.788	(14.04, 23.54)	0.953
	2,500	19.518	(17.22, 21.82)	0.985	19.638	(17.45, 21.83)	0.967
	5,000	19.803	(18.16, 21.44)	0.987	19.833	(18.34, 21.32)	0.980
	10,000	19.941	(18.87, 21.01)	0.977	19.942	(18.94, 20.94)	0.963

as before. For this simulation, we employ the Hoeffing bootstrap for $\mathcal{U}$ statistics once more. Table 6 below presents the results.

The discrete estimators demonstrate finite sample bias. Nonetheless, they also demonstrate convergence toward $\Delta_{1,0}$ as n becomes large, while $\hat{\Delta}_{\text{OLS}}$ remains confounded. Although $\Delta_{1,0} = 5$ is a small to moderate effect size, $\hat{\Delta}_{\hat{Av}}$ shows ample power to detect it under the null hypotheses that $\Delta_{1,0} = 0$ for reasonable sample sizes when τ is set to 30. Traditional levels of acceptable power are only achieved by the $\hat{\Delta}_{\hat{Av}}$ estimator at $n = 10{,}000$ when $\tau = 50$. The increase in variance and the lowered likelihood of observing some values of the support hinder both estimators' performance. Although the mid-range estimator seems to possess less bias for these simulations, the plug-in estimator seems to possess more power.

Our last experiment demonstrates how linear regression can still be used for causal inference when mean exchangeability does not hold, but each error term is a $\mathcal{U}$ variable. To this end, we use a new setup for any measurable function g: $T \sim Bern(0.3)$, $U_1 \sim TN(m = -10, M = 10, \mu = 0, \sigma = 2)$, and $\mu_T = 100 + 20T$. Then we will say $U_2 \sim TN(m = \mu_T - g(T) - 10, M = \mu_T - g(T) + 10, \mu = 0, \sigma = 2)$, $Y|T = \mu_T + U_1$, and $Y(T) = g(T) + U_2$.

For simplicity, we set $g(T) = 90 + 10T$. The underlying model that results from these specifications can also be written as follows: $Y|T = Y(T) + 10 + 10T + \gamma$, where $\gamma = U_1 - U_2$. This model structure can be amended to include more covariates. However, this is unnecessary for our demonstration. What is important is, conditional on a design matrix $\mathbf{x}$, the functional averages are preserved and that linearity holds for $Y|\mathbf{x}$. The true generating process for the counterfactual distribution can be otherwise unknown. Here, $\Delta_{1,0} = 20$. However, confounding is present (in addition to other violations) since $E\{Y(1)\} - E\{Y(0)\} = 10 \neq E(Y|1) - E(Y|0) = 20$.

The results are provided in Table 7. The $\hat{CI}_t$ column provides the arithmetic average of the standard t-distribution confidence set endpoints. For reference, we also include $\hat{CI}_{\mathcal{U}}$ for confidence sets constructed from the concentration inequality in Table 1 for $\mathcal{U}$ errors.

$\hat{CI}_{\mathcal{U}}$ is constructed as follows. Say $\mathbf{x}$ is the design matrix for a regression to estimate $\boldsymbol{\beta}$ and $\mathbf{w} = (\mathbf{x}^\top\mathbf{x})^{-1}\mathbf{x}^\top$. Then $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{w}\boldsymbol{\varepsilon}$. Therefore, $\hat{\beta}_T - \beta_T = \sum_{i=1}^{n} w_{i,s}\varepsilon_i$, where s is the row of $\mathbf{w}$ corresponding to treatment feature. Algebraic rearrangement of the concentration inequality then yields confidence sets of the form $\hat{\beta}_T \pm \sqrt{\sum_{i=1}^{n} R_i^2} \cdot \sqrt{6^{-1}\log(2/\alpha)}$, where R_i is the population range of $w_{i,s}\varepsilon_i$. Under the assumption of a valid mean model specification, the maximum of the supports of the ε_i is feasibly estimable with $\hat{e}_{(n)}$ WLOG, where $\hat{e}$ is a typical residual. Hence, we can use an approximate confidence set of the following form when the extremes of the support of Y are unknown: $\hat{\beta}_T \pm \{\hat{e}_{(n)} - \hat{e}_{(1)}\} \cdot \sqrt{\sum_{i=1}^{n} w_{i,s}^2} \cdot \sqrt{6^{-1}\log(2/\alpha)}$. We contrast these confidence sets to those constructed with the Hoeffding bootstrap of Section 2.

Since the properties of linear regression are well understood, a thorough discussion is unnecessary. The results again substantiate the utility of $\mathcal{U}$ random variables. Under their framework, efficient estimation of causal parameters is more readily achievable, especially if mutual independence is a feasible assumption.

Table 5: Continuous effect estimators: $\Delta_{1,0} = 10$

$\tau = 5$	n	$\hat{\Delta}_{\text{OLS}}$	$\hat{\Delta}_{\text{MR}}$	$\hat{CI}_H$	EP_H
	500	33.45	11.76	(1.53, 22)	0.71
	2,500	34.39	11.69	(3.05, 20.33)	0.85
	5,000	35.09	11.65	(3.67, 19.63)	0.90
	10,000	35.18	11.47	(3.88, 19.07)	0.92
$\tau = 25$	n	$\hat{\Delta}_{\text{OLS}}$	$\hat{\Delta}_{\text{MR}}$	$\hat{CI}_H$	EP_H
	500	33.44	12.82	(10.47, 36.11)	0.05
	2,500	34.40	10.84	(2.73, 18.95)	0.91
	5,000	35.11	10.47	(5.79, 15.15)	1
	10,000	35.20	10.25	(7.68, 12.82)	—

All empirical coverage estimators returned 1 for $\hat{\Delta}_{\text{MR}}$.

Table 6: Discrete effect estimators: $\Delta_{1,0} = 5$

	n	$\hat{\Delta}_{\text{OLS}}$	$\hat{\Delta}_{\hat{Av}}$	$\hat{CI}_H$	$\hat{EP}_H$	$\hat{\Delta}_{\hat{MR}}$	$\hat{CI}_H$	EP_H
$\tau = 30$	500	9.694	6.059	(0.55, 11.57)	0.681	5.899	(–0.47, 12.27)	0.447
	2,500	9.875	5.91	(1.34, 10.48)	0.871	5.861	(0.63, 11.09)	0.680
	5,000	10.017	5.8	(1.45, 10.15)	0.890	5.757	(0.83, 10.69)	0.728
	10,000	10.037	5.765	(1.68, 9.85)	0.921	5.729	(1.19, 10.27)	0.805
$\tau = 50$	500	9.696	6.443	(–0.16, 13.04)	0.458	6.192	(–2.11, 14.49)	0.236
	2,500	9.875	6.268	(0.75, 11.79)	0.728	6.150	(–0.69, 12.99)	0.419
	5,000	10.017	6.147	(0.87, 11.42)	0.734	6.052	(–0.48, 12.59)	0.457
	10,000	10.036	6.074	(1.08, 11.07)	0.804	5.995	(–0.03, 12.02)	0.551

All empirical coverage estimators were > 0.998.

Table 7: Linear regression for functional averages

n	$\hat{\Delta}_{\text{OLS}}$	$\hat{CI}_t$	$\hat{CI}_{\mathcal{U}}$	$\hat{CI}_H$
500	20	(19.62, 20.37)	(19.09, 20.91)	(19.9, 21.58)
2,500	—	(19.83, 20.17)	(19.52, 20.48)	(19.28, 20.71)
5,000	—	(19.88, 20.12)	(19.64, 20.35)	(19.49, 20.51)
10,000	—	(19.91, 20.09)	(19.74, 20.26)	(19.64, 20.36)

5 A data application

In this section, we employ NHEFS data to demonstrate our concepts. The NHEFS conducted medical examinations from 1971–1975 from noninstitutionalized civilian adults aged 24–74 (N = 14,407) in the United States as part of a national probability sample. Follow-up surveys were then administered in 1982, 1984, and subsequent years to collect measurements for behavioral, nutritional, and clinical variables. Further documentation is available elsewhere [58]. The subset of data we use here (n = 1,479) originates from the original 1971 medical examination and follow-up in 1982.

Exercise (0: moderate to much; 1: little to none) is the treatment variable (T) of interest. Age, chronic bronchitis/emphysema diagnosis (1: yes; 0: never), education attained in 1971 (1: < 8th grade; 2: HS dropout; 3: HS; 4: college dropout; 5: college), income, race (1: non-white; 0: white), sex (1: female; 0: male), years smoking, alcohol frequency, and weight (kilograms) are utilized as adjusting covariates (henceforth denoted as $\mathbf{L}$). Although SBP was measured with integer values, it is still treated as continuous in most instances. Pertinently, T and most covariates were all measured in 1971. Only SBP and weight were measured in 1982. All continuous covariates are centered on their observed sample means for this analysis.

Here, we are interested in seeing if a history of exercise exerted a causal effect on SBP in smokers. We aim to estimate $E\{Av\{Y(1)|\mathbf{L}\} - Av\{Y(0)|\mathbf{L}\}\}$ and $E\{Av\{Y(1)|S\} - Av\{Y(0)|S\}\}$ as summary causal effects, where $S = s$ represents a stratum that has been constructed from the quintiles of $\hat{\mathbf{e}}(L)$, the estimated propensity scores. Recall that these estimands represent causal effects that have been averaged over all strata, either probabilistically or uniformly. Since we are presupposing linear models, these estimands are simply the treatment coefficients of each respective specification. Our goal is also to estimate the marginal functional average causal effect, $Av\{Y(1)\} - Av\{Y(0)\}$. Although we do not target expected causal effects intentionally, recall that, under the suppositions of Proposition 5, if C1-C3 do hold in addition, then we are implicitly targeting expected causal effects as well since this setup implies that $Av\{Y(T)|\mathbf{L}\} = E\{Y(T)|\mathbf{L}\}$ and hence that $Y(T)|\mathbf{L} \in \mathcal{U}$.

For this reason, and because attempting to account for the salient forces determining a phenomenon is good practice scientifically, we still try to achieve C1-C3 in this analysis. To this end, race is considered a confounder since it represents both genetic information and socio-historical constructions [59]. In a similar vein, we choose to adjust for sex to account for possible biological influences, and since it is also an imperfect

proxy for social institutions that can impact exercise habits, other health behaviors, and therefore blood pressure. All other variates mentioned are adjusted for since they are either known to affect both SBP and exercise habits directly or to act as conduits for more general institutional or ecological influences that can transcend time. Theoretically, adjusting for them can help to block backdoor paths from a subset of unknown confounders, which, although mostly irrelevant here in terms of their probabilistic effects insofar as functional averages are concerned, can still impact supports. Furthermore, it seems intuitive to assert that fewer omitted variables can translate into a heightened likelihood that the probabilistic errors remaining follow a $\mathcal{U}$ pattern, although this is not guaranteed.

5.1 Methods

The following models are employed to estimate possible causal effects: simple linear regression (SLR) to estimate $Av\{Y(1)\} - Av\{Y(0)\}$ under the working supposition that each outcome distribution is $\mathcal{U}$ class, multiple linear regression (MLR) to estimate $E\{Av\{Y(1)|\mathbf{L}\} - Av\{Y(0)|\mathbf{L}\}\}$, and linear regression adjusted for propensity score strata (PS) to estimate $E\{Av\{Y(1)|S\} - Av\{Y(0)|S\}\}$. For contrast, we also estimate $Av\{Y(1)\} - Av\{Y(0)\}$ with the discrete plug-in (Av) from Section 2.1. In a similar vein, we also estimate $E\{Av\{Y(1)|\mathbf{L}\} - Av\{Y(0)|\mathbf{L}\}\}$ with the AIPW estimator defined in Section 3. This is accomplished using the multiple linear regression model, an application of the parametric g-formula, and estimated propensity scores. Propensity scores are estimated with logistic regression using the same covariates as the regression model. No covariate transformations are used for the logit model, although we introduce higher-order terms to the regression if it appears to improve linearity.

Standard t-statistic-based confidence sets are employed for the coefficient estimators of the regression models. We use the Hoeffding bootstrap of Section 3 for equation (2.1) plug-in and a standard bootstrapping procedure for the AIPW estimator, both with $B = 1{,}000$ replications. The null hypothesis that $E\{Av\{Y(1)|\mathbf{L}\} - Av\{Y(0)|\mathbf{L}\}\} = E\{Y(1)\} - E\{Y(0)\}$ is also tested via a standard bootstrapping procedure of the difference between the AIPW and MLR estimators. All tests are conducted at the $\alpha = 0.05$ level using R version 4.2.2 statistical software [60]. The assumptions of $\mathcal{U}$ status and valid mean model specification are substantiated by inspecting residual versus fitted plots and empirical CDF plots.

5.2 Results

The coefficient estimate for the simple linear regression is $\hat{\beta}_t = -3.87$ (95% CI: −5.821, −1.916), while the plug-in functional average estimate for $Av\{Y(1)\} - Av\{Y(0)\}$ is −3.81 (95% CI: -24.17, 17.42). The latter estimates that, on average, the value of an adult smoker's SBP is 3.81 mmHg less when they possess a history of exercise, although this estimate was not statistically significant. Additional model results are available in Table 8.

Initial fitting procedures for the multiple linear regression model showed mild departures from linearity. The addition of a quadratic term for age appeared to improve model fit. From this reformed model, the estimate for the expected functional average change in SBP is $\hat{E}\{Av\{Y(1)|\mathbf{L}\} - Av\{Y(0)|\mathbf{L}\}\} = -0.736$ (95% CI: −2.56, 1.09), meaning that adult smokers with a history of exercise were measured to have 0.736 mmHg less SBP on average than those who did not exercise, with all other covariates held equal. This estimate is not statistically significantly different from the AIPW estimate {Difference = −0.043 (95% CI : −0.414, 0.329)}, which

Table 8: Summary of model results

Parameter	Method	Estimate	95% Confidence interval
$Av\{Y(1)\}$–$Av\{Y(0)\}$	Av	–3.81	[–24.17, 17.42]
$E\{Av\{Y(1)\|\mathbf{L}\}$–$Av\{Y(0)\|\mathbf{L}\}\}$	MLR	–0.736	[–2.56, 1.09]
$E\{Av\{Y(1)\|\mathbf{L}\}$–$Av\{Y(0)\|\mathbf{L}\}\}$	AIPW	–0.69	[–2.47, 1.09]
$E\{Av\{Y(1)\|S\}$–$Av\{Y(0)\|S\}\}$	PS	–1.014	[–2.96, 0.94]

estimates that the SBP of adult smokers is 0.69 less on average when they report a history of exercise (95% CI: −2.47, 1.09). Finally, the estimate of the expected functional average change in SBP w.r.t. the propensity score stratified model is $\hat{E}\{Av\{Y(1)|S\} - Av\{Y(0)|S\}\} = -1.014$ (95% CI: −2.96, 0.94).

5.3 Model checking

Empirical CDF plots for the SBP distributions are presented by propensity score strata and by treatment status in Figure 1. The residual versus fitted plot for the MR model is also included.

We note no remarkable departures from linearity in the residual versus fitted plot for the multiple linear regression model. Moreover, the residuals appear to possess an approximately symmetric spread around the zero horizontal line, although some outlying points appear to violate the critical assumption defining $\mathcal{U}$ status in this context: that the errors are supported around symmetric extremes. Altogether, although departures from strict $\mathcal{U}$ status are observable, the assumption of $\mathcal{U}$ status appears to be feasibly met. The stratified empirical CDF plots in Figure 1 do not appear to corroborate approximate sum-symmetric behavior since they show more area below the functional lines than above in most cases. This is also true for the empirical CDF plots of the nonadjusted conditional distributions. Hence, the results of the simple t-test cannot be afforded a causal interpretation w.r.t. a change in functional average.

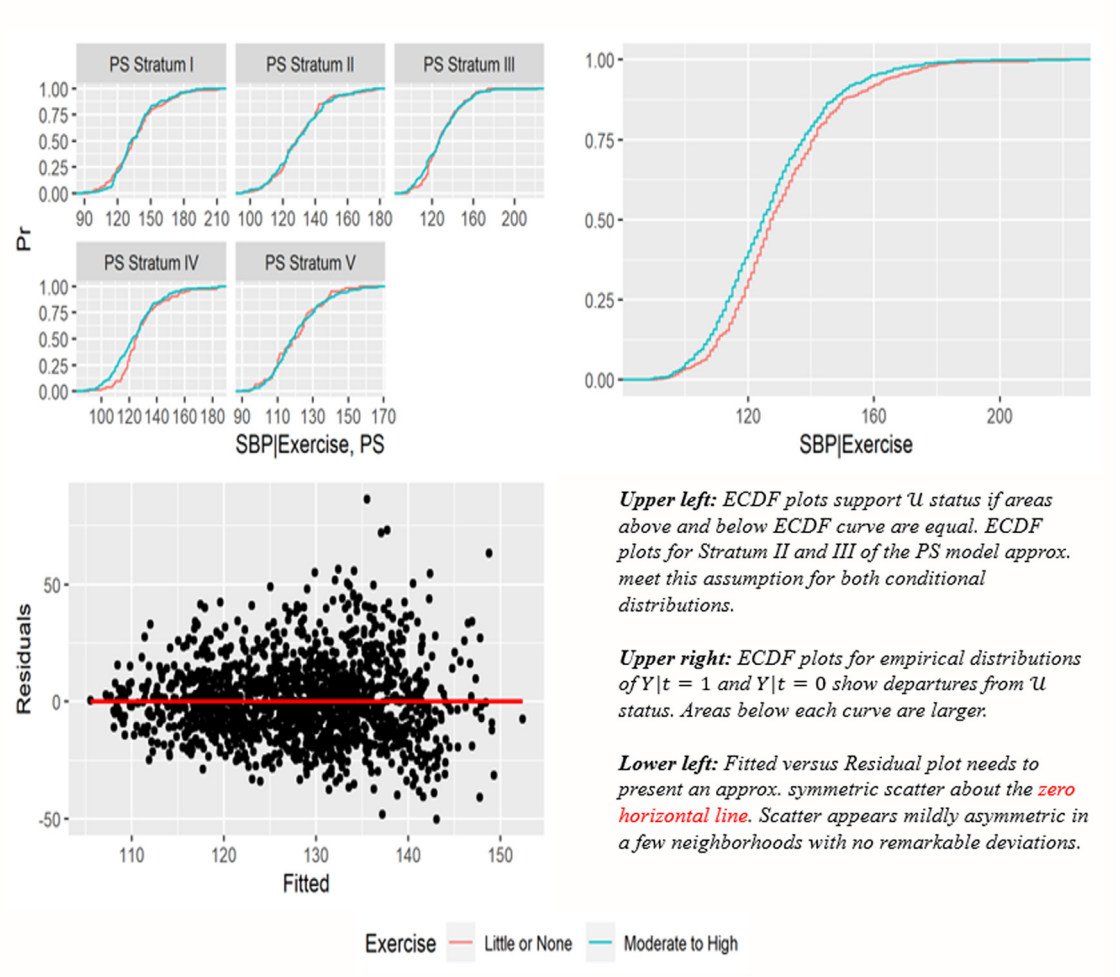

Figure 1: Model validation plots.

5.4 Discussion

Since the NHEFS was a national probability sample of noninstitutionalized adults, there is little reason to believe that A4 was not fulfilled, conditional on our adjusting covariates. Recall that positing the opposing notion in this context is to affirm – for noninstitutionalized adults who smoked – that there were possible values of SBP in each treatment population that had zero probability of being observed in the sample of the observational design. Insofar as the NHEFS survey was truly a probability sample and hence noninformative, rejecting A4 also means that a subset of potential SBP values could never have been observed in the real world, in all likelihood. Stating this also means that there did not exist a set of adjusting variables s.t. expected causal effects could have been identified and estimated. Due to the feasible validity of A4 and the fact that $\mathcal{U}$ status appeared to be approximately verified for the basic multiple regression model, we are confident that – conditional on our covariates – the expected (functional average) causal effects were sufficiently identified and that the average effect of a history of exercise upon SBP in adult smokers is plausibly within a neighborhood of zero. This conclusion is further corroborated by the unadjusted estimate of the difference in functional averages between treatment groups, which was also not statistically significantly different from zero. However, this test was hampered by the fact that each conditional SBP population possessed a relatively light right tail and the sample size was small. In these circumstances, the mid-range and the discrete plug-in estimator are inefficient statistics. This fact, combined with the intrinsic conservatism of the Hoeffding bootstrap procedure, produced a confidence set with a length that was substantially greater than the lengths of others.

Vitally, we have little reason to believe that we successfully adjusted for all confounding variables. Hence, we do not purport to interpret the effect estimates as expected treatment effects. Nevertheless, since the AIPW estimator was not statistically significantly different from its functional average counterpart, and the outcome model seemed reasonably well specified, this at least seemed to provide evidence that bias was fairly attenuated. Of course, if mean exchangeability and consistency did hold, they would also be estimates of this type of contrast.

It is also apropos to note that the consistency assumption might be violated in this analysis. This is because respondents were asked if they exercised little to none, moderately, or much; however, the meanings of these words possess no absolutism. Hence, it is possible that multiple exercise treatments existed under the premise of one coding. This does not undermine what is formally specified in A4, at least conditional on the variables observed, although it does complicate the generalizability of the results if present.

6 Conclusion

We demonstrated that causal inference is achievable in the absence of mean exchangeability if the support of the counterfactual distribution is preserved. Moreover, we offered exposition on the possible utility and scientific meaningfulness of functional average change. To overcome some of the difficulties of functional average estimation, we introduced a simple class of random variables – the $\mathcal{U}$ class – that possesses a milieu of practical properties. By using the $\mathcal{U}$ random variable framework, we showed that ubiquitously employed statistical procedures produce estimates with causal interpretations under exceptionally mild conditions, many of which are already supposed in most applied settings to investigate associations. Hence, even if a researcher fails to control for all confounding variables, she still might be left with a second-prize of sorts, and one that possesses salient causal meaning. Since uncontrolled confounding is safely assumed to be nearly omnipresent outside of toy examples, we believe that this framework provides a strong defense of elementary methods. Further work is of course due. For instance, this article did not explore how sensitive the estimation of functional averages is to departures from $\mathcal{U}$ status in the regression errors. This is a critical question. Moreover, we observe that we did not pursue the sample minimum, maximum, or any other function of the supports as causal estimators in and of themselves, although the set of assumptions employed here also establishes their utility in this regard. Developing this area of theory will most certainly be advantageous.

Much of extreme value theory can be immediately transformed into theory for causal inference, for instance. Furthermore, we restricted much of our analyses to basic linear models. However, the results of this article apply just as equally to nonlinear models, semi-parametric models, or causal estimands estimated via machine learning algorithms more generally.

Finally, we also presented a new approach to the bootstrapping process. We called this approach the Hoeffding bootstrap. Ultimately, we showed that a simple pair of inscrutable inequalities can be wielded to construct feasible and mostly conservative confidence sets for a very general class of statistics. Pertinently, although we excluded some dependency scenarios, they were only of the most extreme type. We cited no particular dependency theory, otherwise. Nevertheless, much of our justification for this procedure was asymptotic. This is a good start. However, this cannot be the end. A defensible set of sufficient conditions that ensure the approach for small samples will do much to lighten the burden of uncertainty. We conjecture that a body of literature on the behavior of extreme order statistics and their expected values, but within the universe of bounded random variables, will do much to improve this method.

Acknowledgements: We thank Dr. Miguel Hernán for making the NHEFS data accessible. Furthermore, we would like to thank the editor and reviewers for their diligent and insightful commentary, which substantially improved the quality of this manuscript.

Funding information: We have no funding information to declare.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and consented to its submission to the journal, reviewed all the results, and approved the final version. SS developed the theoretical content, proofs, and simulation experiments and prepared the manuscript. EG supervised the preparation of the manuscript, reviewed its contents, contributed to the data application, and participated in manuscript editing. LZ also supervised the manuscript preparation and participated in editing. Additionally, LZ reviewed and commented on the manuscript's proofs and contributed to its mathematical content.

Conflict of interest: The authors state no conflicts of interest.

Data availability statement: The NHEFS data were acquired from Dr. Miguel Hernán's faculty website (https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/), where it is publically accessible.

References

[1] Rubin DB. Essential concepts of causal inference: a remarkable history and an intriguing future. Biostat Epidemiol. 2019;3(1):140–55.
[2] Pearl J. Causal inference. Causality: objectives and assessment. Cambridge, MA, USA: MIT Press; 2010. p. 39–58.
[3] Holland PW. Statistics and causal inference. J Amer Stat Assoc. 1986;81(396):945–60.
[4] Imbens GW, Rubin DB. Causal inference in statistics, social, and biomedical sciences. Cambridge: Cambridge University Press; 2015.
[5] Ding P, Li F. Causal inference. Stat Sci. 2018;33(2):214–37.
[6] Hernán MA, Robins JM. Causal inference. Boca Raton, FL: CRC; 2010.
[7] Imbens GW, Rubin DB. Rubin causal model. In: Microeconometrics. London: Palgrave Macmillan; 2010. p. 229–41.
[8] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.
[9] Holland PW, Rubin DB. Causal inference in retrospective studies. ETS Res Report Series. 1987;1987(1):203–31.
[10] Jin H, Rubin DB. Principal stratification for causal inference with extended partial compliance. J Amer Stat Assoc. 2008;103(481):101–11.
[11] Belloni A, Chernozhukov V, Fernandez-Val I, Hansen C. Program evaluation and causal inference with high-dimensional data. Econometrica. 2017;85(1):233–98.
[12] Gangl M. Causal inference in sociological research. Ann Rev Soc. 2010;36:21–47.

[13] Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? Epidemiology. 2009;20(1):3–5.
[14] Hernán MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Commun Health. 2006;60(7):578–86.
[15] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10:37–48.
[16] Greenland S, Pearl J, Robins JM. Confounding and collapsibility in causal inference. Stat Sci. 1999;14(1):29–46.
[17] Glantz SA, Parmley WW. Passive smoking and heart disease. Epidemiology, physiology, and biochemistry. Circulation. 1991;83(1):1–12.
[18] Stallones RA. The association between tobacco smoking and coronary heart disease. Int J Epidemiol. 2015;44(3):735–43.
[19] Narkiewicz K, Kjeldsen SE, Hedner T. Is smoking a causative factor of hypertension? Blood Pressure. 2005;14(2):69–71.
[20] Elley CR, Arroll B. aerobic exercise reduces systolic and diastolic blood pressure in adults. Evidence Based Med. 2002;7(6):170.
[21] Park W, Miyachi M, Tanaka H. Does aerobic exercise mitigate the effects of cigarette smoking on arterial stiffness? J Clin Hypertension. 2014;16(9):640–4.
[22] Pfeffermann D, Sverchkov M. Inference under informative sampling. In: Handbook of statistics. Vol. 29. Elsevier; 2009. p. 455–87.
[23] Pfeffermann D, Krieger AM, Rinott Y. Parametric distributions of complex survey data under informative probability sampling. Statistica Sinica. 1998;8:1087–114.
[24] Patil GP, Rao CR, Zelen M, Patil GP. Weighted distributions. Wiley, New York: Citeseer; 1987.
[25] Patil GP, Rao CR. Weighted distributions and size-biased sampling with applications to wildlife populations and human families. Biometrics. 1978;38:179–89.
[26] Pearl J. Statistics and causal inference: A review. Test. 2003;12:281–345.
[27] Hansen BE. Uniform convergence rates for kernel estimation with dependent data. Econ Theory. 2008;24(3):726–48.
[28] Chen YC. A tutorial on kernel density estimation and recent advances. Biostat Epidemiol. 2017;1(1):161–87.
[29] Zambom AZ, Ronaldo D. A review of kernel density estimation with applications to econometrics. Int Econ Rev. 2013;5(1):20–42.
[30] Chernoff H, Gastwirth JL, Johns MV. Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. Ann Math Stat. 1967;38(1):52–72.
[31] Hosking JR. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. J R Stat Soc Ser B (Methodological). 1990;52(1):105–24.
[32] Bickel PJ. On some analogues to linear combinations of order statistics in the linear model. Ann Stat. 1973;1:597–616.
[33] David HA, Nagaraja HN. Order statistics. Hoboken, NJ, USA: John Wiley & Sons; 2004.
[34] Barndorff-Nielsen O. On the limit behaviour of extreme order statistics. Ann Math Stat. 1963;34(3):992–1002.
[35] Sparkes S, Zhang L.. Properties and deviations of random sums of densely dependent random variables; 2023. https://arxiv.org/abs/2310.11554.
[36] Bingham N. The sample mid-range and symmetrized extremal laws. Stat Probabil Lett. 1995;23(3):281–8.
[37] Bingham N. The sample mid-range and interquartiles. Stat Probabil Lett. 1996;27(2):131–6.
[38] Broffitt JD. An example of the large sample behavior of the midrange. Amer Stat. 1974;28(2):69–70.
[39] Arce GR, Fontana SA. On the midrange estimator. IEEE Trans Acoustics Speech Signal Proces. 1988;36(6):920–2.
[40] Efron B, Tibshirani RJ. An introduction to the bootstrap. United States: CRC Press; 1994.
[41] Bickel PJ, Freedman DA. Some asymptotic theory for the bootstrap. Ann Stat. 1981;9(6):1196–217.
[42] Bickel PJ, Sakov A. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. Statistica Sinica. 2008;18:967–85.
[43] Bickel PJ, Ren JJ. The bootstrap in hypothesis testing. Lecture Notes-Monograph Series. 2001:91–112.
[44] Swanepoel JW. A note on proving that the (modified) bootstrap works. Commun Stat-Theory Methods. 1986;15(11):3193–203.
[45] Beran R, Ducharme GR. Asympotic theory for bootstrap methods in statistics. Montréal, Québec: Les Publications CRM; 1991.
[46] Politis DN, Romano JP, Wolf M. On the asymptotic theory of subsampling. Statistica Sinica. 2001;11:1105–24.
[47] Shao X. The dependent wild bootstrap. J Amer Stat Assoc. 2010;105(489):218–35.
[48] Hall P, Horowitz JL, Jing BY. On blocking rules for the bootstrap with dependent data. Biometrika. 1995;82(3):561–74.
[49] Kreiss JP, Paparoditis E. Bootstrap methods for dependent data: A review. J Korean Stat Soc. 2011;40(4):357–78.
[50] Lahiri SN. Resampling methods for dependent data. New York: Springer Science & Business Media; 2003.
[51] Rider PR. The midrange of a sample as an estimator of the population midrange. J Amer Stat Assoc. 1957;52(280):537–42.
[52] Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: A review. Rev Econ Stat. 2004;86(1):4–29.
[53] Mansournia MA, Altman DG. Inverse probability weighting. BMJ. 2016;352.
[54] Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. Amer J Epidemiol. 2011;173(7):761–7.
[55] Kang JD, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci. 2011;22(4):523–39.
[56] Glynn AN, Quinn KM. An introduction to the augmented inverse propensity weighted estimator. Politic Anal. 2010;18(1):36–56.
[57] Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. Int J Epidemiol. 2017;46(2):756–62.
[58] Madans JH, Kleinman JC, Cox CS, Barbano HE, Feldman JJ, Cohen B, et al. 10 years after NHANES I: report of initial followup, 1982–84. Public Health Reports. 1986;101(5):465.
[59] Witzig R. The medicalization of race: scientific legitimization of a flawed social construct. Ann Internal Med. 1996;125(8):675–9.
[60] R Core Team. R: A language and environment for statistical computing. Vienna, Austria; 2022. Available from: https://www.R-project.org/.