Causal Inference Under Outcome-Based Sampling with Monotonicity Assumptions

Sung Jae Jun\textsuperscript{a} and Sokbae Lee\textsuperscript{b}

\textsuperscript{a}Department of Economics, Pennsylvania State University, University Park, PA; \textsuperscript{b}Department of Economics, Columbia University, New York, NY

\textbf{ABSTRACT}

We study causal inference under case-control and case-population sampling. Specifically, we focus on the binary-outcome and binary-treatment case, where the parameters of interest are causal relative and attributable risks defined via the potential outcome framework. It is shown that strong ignorability is not always as powerful as it is under random sampling and that certain monotonicity assumptions yield comparable results in terms of sharp identified intervals. Specifically, the usual odds ratio is shown to be a sharp identified upper bound on causal relative risk under the monotone treatment response and monotone treatment selection assumptions. We offer algorithms for inference on the causal parameters that are aggregated over the true population distribution of the covariates. We show the usefulness of our approach by studying three empirical examples: the benefit of attending private school for entering a prestigious university in Pakistan; the relationship between staying in school and getting involved with drug-trafficking gangs in Brazil; and the link between physicians’ hours and size of the group practice in the United States.

1. Introduction

Random sampling is convenient for causal inference, but it may be too costly in practice for various reasons. For instance, rare events are likely to be severely under-represented in a random sample of a finite size: for example, cancer (Breslow and Day 1980), infant death (Currie and Neidell 2005), consumer bankruptcy (Domowitz and Sartain 1999), entering a highly prestigious university (Delavande and Zafar 2019), and drug trafficking (Carvalho and Soares 2016). The objective of this article is to study causal inference in outcome-based sampling scenarios such as case-control or case-population studies.

We focus on observational data, as opposed to experimental ones, with a binary outcome and a binary treatment. Holland and Rubin (1988) adopt the potential outcome framework to show that the assumption of strong ignorability can be used to identify the counterfactual odds ratio in case-control studies. They then argue that the counterfactual odds ratio approximates the ratio of two potential-outcome probabilities (i.e., causal relative risk) under the rare-disease assumption, which says that the probability of outcome occurrence (e.g., having "a certain disease") is close to zero. Their work is our starting point, and we make additional contributions in several ways.

First, we focus on two direct causal parameters (i.e., a ratio or a difference of two potential-outcome probabilities) that are more straightforward to interpret than counterfactual odds ratios: our parameters of interest are the causal relative and attributable risks given a specific value of covariates. Second, we do not appeal to the rare-disease assumption, and we take the perspective of partial identification (see, e.g., Manski 1995, 2003, 2007; Tamer 2010; Molinari 2020, among others). Third, we consider a set of monotonicity assumptions, and we compare their identification power with that of strong ignorability. Strong ignorability is a popular setup for causal inference, but its identification power in outcome-based sampling turns out to be somewhat limited. Specifically, in case-control or case-population studies, strong ignorability is generally not sufficient to point identify the causal relative and attributable risks. We can obtain bounds on them, but they are not much better than those we can obtain in a less restrictive setup using monotonicity. Specifically, we will consider monotone treatment response (Manski 1997, MTR hereafter) and monotone treatment selection (Manski and Pepper 2000, MTS hereafter).

Our work builds upon Manski (2007, chap 6), who conducts a partial identification analysis for both relative and attributable risks under outcome-based sampling without focusing on causal parameters. The MTR and MTS assumptions as well as other related notions of monotonicity have been extensively used in the literature. For example, see Vythacil and Yildiz (2007), Bhattacharya, Shaikh, and Vythacil (2008, 2012), Pearl (2009), VanderWeele and Robins (2009), Kreider et al. (2012), Jiang, Chiba, and VanderWeele (2014), Okumura and Usui (2014), Choi (2017), Kim et al. (2018), and Machado, Shaikh, and Vytlacil (2019) among others.

We now discuss the relation of our work with the existing literature on causal inference under outcome-based sampling. Månnsson et al. (2007) point out that the propensity score method has only limited ability to control for confounding factors in case-control studies. Our method does not rely on the propensity score. Rose (2011) and Van der Laan and Rose (2011) use an assumption that the true case probability is known by a prior
study. We focus on the instance of unknown case probability. Didelez and Evans (2018) provide an extensive survey on causal inference in case-control studies, but no discussion on partial identification approaches can be found there. Therefore, possibilities based on partial identification appear to be rather under-explored. Kuroki, Cai, and Geng (2010) and Gabriel, Sachs, and Sjölander (2022) are notable exceptions. Gabriel, Sachs, and Sjölander (2022) obtain bounds on the causal attributable risk in a variety of scenarios including outcome-dependent sampling with an instrumental variable. But they do not leverage any monotonicity assumption, while we do not consider instrumental variables but we use monotonicity restrictions. Kuroki, Cai, and Geng (2010) is more similar to our work in that they obtain bounds on both the causal relative and attributable risks by using the MTR assumption. Our contributions relative to Kuroki, Cai, and Geng (2010) can be highlighted as follows: (a) we exploit not only the MTR but also the MTS assumption, and therefore the bounds are different; (b) we consider case-control sampling as well as case-population sampling; (c) we compare the identification power of the popular assumption of strong ignorability with that of the MTR and MTS assumptions; (d) we consider how to aggregate the causal parameters over the distribution of the covariates; and (e) we provide algorithms for causal inference.

The remaining part of the article is organized as follows. In Section 2 we formally present the setup including the causal parameters of interest and the sampling schemes. Sections 3 and 4 focus on the causal relative risk and attributable risk to address identification and aggregation. Section 5 covers how to carry out causal inference. Section 6 presents three empirical applications. Specifically, by using datasets collected in previous studies, we address new research questions that are not examined in the original papers. All the proofs, discussions on semiparametric efficiency and computational algorithms are in Online Appendix. An accompanying R package is available on the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=ciccr, and the replication files are available at https://github.com/sokbae/replication-JunLee-JBES.

2. Preliminaries

2.1. Causal Parameters

Let \((Y^*, T^*, X^*)\) be a random vector of a binary outcome, a binary treatment, and covariates of a representative individual. Since we are interested in outcome-based sampling, we assume that a random sample of \((Y^*, T^*, X^*)\) is not available. Instead, we have a sample of \((Y, T, X)\), where the distribution of \((T, X)\) given \(Y\) is related with that of \((T^*, X^*)\) given \(Y^*\). The exact sampling schemes and related assumptions will be discussed in detail later, and in this section, we only focus on the parameters of interest.

For the sake of causal inference, we use the usual potential outcome notation. So, \(Y^*(t)\) will be the potential outcome for \(T^* = t\), and \(Y^*\) can be written as \(Y^* = Y^*(1)T^* + Y^*(0)(1 − T^*)\). Therefore, our notation extends Chen (2001) and Xie et al. (2020) by adding an extra layer of potential outcomes. The causal effect of the treatment can be measured by either (conditional) relative risk or attributable risk: each of them is defined as follows:

\[
\theta_{RR}(x) := \frac{\mathbb{P}\{Y^*(1) = 1|X^* = x\}}{\mathbb{P}\{Y^*(0) = 1|X^* = x\}},
\]

\[
\theta_{AR}(x) := \mathbb{P}\{Y^*(1) = 1|X^* = x\} - \mathbb{P}\{Y^*(0) = 1|X^* = x\},
\]

provided that the denominator of \(\theta_{RR}(x)\) is strictly positive. Therefore, \(\theta_{AR}(x)\) is the usual conditional average treatment effect, whereas \(\theta_{RR}(x)\) is a causal version of the relative risk parameter.

Relative risk defined by a ratio of “success” probabilities has been popular in epidemiology and biostatistics, particularly when the “success” is a rare event: if the treatment changes the success probability from 0.01 to 0.02, then it is a 100% increase, though the difference of 0.01 may suggest an impression that the change was unimportant. Further, it turns out that \(\theta_{RR}(x)\) is closely related with the odds ratio (in terms of the observed variables), which has been widely used as a measure of association in case-control studies.

2.2. Bernoulli Sampling

As we mentioned earlier, we assume that a random sample of \((Y^*, T^*, X^*)\) is unavailable. Instead the researcher has access to a random sample of \((Y, T, X)\), where the distribution of \((Y, T, X)\) is related with that of \((Y^*, T^*, X^*)\) by Bernoulli sampling (e.g., Breslow, Robins, and Wellner 2000) that we describe below.

In Bernoulli sampling, the researcher first draws a Bernoulli variable \(Y\) from a pre-specified marginal distribution, after which she randomly draws \((T, X)\) from some \(P_y\) if and only if \(Y = y\); so, \(Y\) is an artificial device to decide which subpopulation we will draw \((T, X)\) from. If \(P_y\) is the distribution of \((T^*, X^*)\) conditional on \(Y^* = y\), then this is nothing but case-control sampling. Since \(h_0 = \mathbb{P}\{Y = 1\} \in (0, 1)\) is part of the sampling scheme, we will assume that it is known; if not, it can be easily estimated without compromising inferential validity. See online appendices B.1 and B.2 for more details. Before we proceed, we make a common support assumption for simplification.

**Assumption A (Common Support).** The support of \(X^*\) and that of \(X\) given \(Y = y\) for \(y = 0, 1\) coincide; the common support will be denoted by \(\mathcal{X}\).

Below we discuss two leading cases of Bernoulli sampling that we focus on throughout the article.

**Design 1 (Case-Control Sampling).** For \(y \in \{0, 1\}, P_y\) is the distribution of \((T^*, X^*)\) given \(Y^* = y\).

**Design 2 (Case-Population Sampling).** \(P_1\) is the conditional distribution of \((T^*, X^*)\) given \(Y^* = 1\), whereas \(P_0\) represents the distribution of \((T^*, X^*)\) of the entire population.

**Design 1** is arguably the most popular form of case-control studies (e.g., Breslow 1996) and **Design 2** was referred to as “contaminated case-control studies” by Lancaster and Imbens (1996): we call the latter design case-population sampling, which is more descriptive. The case-population sampling design has
been used to study drug trafficking (Carvalho and Soares 2016) and mass demonstrations (Rosenfeld 2017) among others.

Note that the distribution of $(T, X)$ is identified from the data, but that of $(T^*, X^*)$ may not. For instance, in Design 1, we have $f_X(x) = f_{X|Y}(x|1)h_0 + f_{X|Y}(x|0)(1 - h_0) \neq f_{X^*}(x)$, unless $h_0$ is the same as $p_0 := P(Y^* = 1)$, that is, the true probability of the case in the population. Further, $f_{X,T}(x, t) = f_{X|Y}(x|1)h_0 = f_{X^*}(x)P(Y^* = 1|X^* = x)h_0/p_0$, which yields the likelihood function studied in for example Manski and Lerman (1977). We emphasize that $P(Y = 1|X = x)$ does not have an economic interpretation like $P(Y^* = 1|X^* = x)$, where the latter is often specified via domain knowledge in a specific field such as a utility function with an additively separable normal or Gumbel error term.

### 3. Identification

In this section, we study identification of the causal parameters, that is, $\theta_{RR}(x)$ and $\theta_{AR}(x)$. Aggregation over $x \in X$ will be considered later. For the purpose of the identification analysis, we consider two sets of assumptions: one is the standard case of strong ignorability, and the other is an alternative possibility based on monotonicity assumptions. We will see that even strong ignorability is not sufficient to point-identify $\theta_{RR}(x)$ or $\theta_{AR}(x)$ under case-control sampling, that is, Design 1.

We consider the following assumptions.

**Assumption B (Overlap).** For all $(y, t, s, x) \in \{0, 1\}^3 \times X$, $0 < P(Y = y) = P(T = s) = P(X = x) < 1$.

**Assumption C (Unconfoundedness).** For all $(t, x) \in \{0, 1\} \times X$, $P(Y = 1|T = t, X = x) = P(Y = 1|T = 0, X = x)$.

Assumptions B and C together constitute strong ignorability, which is a standard setup for causal inference. Assumption B is stated in terms of the joint probability mass function of $Y^*(t)$ and $T^*$ given $X^* = x$. We do this for a few reasons. First, Assumption B ensures that all the conditional probabilities we consider and their ratios are well-defined: for example, $\theta_{RR}(x)$ is well-defined under Assumption B. Also, it ensures that the distribution of $(Y, T, X)$ has enough overlap to identify $P(T = t|Y = y, X = x)$ under each of the two Bernoulli sampling schemes.

The key component of the strong ignorability setup is Assumption C. In the following sections we will start from clarifying how far Assumption C can take us to identify the causal parameters under case-control and case-population sampling. Although it is standard, strong ignorability does not allow the treatment assignment to be endogenous. Therefore, we consider a set of alternative assumptions under which we study how much we can say about the causal parameters under the two sampling scenarios.

**Assumption D (Monotone Treatment Response).** $Y^*(1) \geq Y^*(0)$ almost surely.

**Assumption E (Monotone Treatment Selection).** For all $t \in \{0, 1\}$ and $x \in X$, $P(Y^*(t) = 1|T^* = 1, X^* = x) \geq P(Y^*(t) = 1|T^* = 0, X^* = x)$.

Assumption D was first proposed by Manski (1997), while Assumption E was used by Manski and Pepper (2000). Assumption D says that treatment is potentially beneficial but it never hurts. For instance, if an individual does not earn high income with a college degree, then the person will not be highly paid without a college degree, either. Assumption E states that, all else being equal, individuals with a higher degree are as likely to earn high incomes if their educational attainment were randomly assigned, as compared to those without a higher degree.

In essence, the treatment decision made by an individual reveals their “type”: continuing with the same example, those opting for a higher degree are more motivated, and they would be at least as likely to earn high incomes as those who choose not to pursue a higher degree if they were randomly assigned to different educational attainment. Assumption E is trivially weaker than Assumption C, and it allows individuals with “higher ability” to self-select a higher degree.

Before we move on, we define the following functions:

$$r_{CC}(x, p) := \frac{p(1 - h_0)P(Y = 1|X = x) - h_0(1 - p)P(Y = 0|X = x)}{h_0P(Y = 1|X = x)} + h_0(1 - p)P(Y = 0|X = x),$$

(3)

$$r_{CP}(x, p) := \frac{p(1 - h_0)P(Y = 1|X = x)}{h_0P(Y = 0|X = x)},$$

(4)

where $P(Y = 1|X = x)$ is the prospective regression function identified from the data. Here, both $r_{CC}(x, p)$ and $r_{CP}(x, p)$ can be alternatively expressed by using the conditional densities of $X$ given $Y = y$ by the Bayes rule, which is related with the distribution of $X^*$ given $Y^* = y$ or simply the distribution of $X^*$, depending on the sampling design. Indeed, it can be shown that $r_{CC}(x, p_0) = P(Y^* = 1|X^* = x)$ under case-control sampling and $r_{CP}(x, p_0) = P(Y^* = 1|X^* = x)$ under case-population sampling, where $p_0 = P(Y^* = 1)$; see lemma A.3 in the online appendix.

Therefore, one can view the functions $r_{CC}$ and $r_{CP}$ as devices to exploit the fact that the only unidentified object in our context will be $p_0$.

#### 3.1. Causal Relative Risk

In this section, we study identification of $\theta_{RR}(x)$, for which we first introduce some notation. Let $P(t|y, x) = P(T = t|Y = y, X = x)$ be the retrospective regression function. For $(x, p) \in X \times \{0, 1\}$ and for $d \in \{CC, CP\}$, define

$$\Gamma_{d,RR}(x, p) := \frac{\Pi(1|1, x)}{\Pi(0|1, x)} \times \frac{\Pi(0, 0, x) + r_{d}(x, p)\Pi(0|1, x) - \Pi(0, 0, x)}{\Pi(0, 0, x) + r_{d}(x, p)\Pi(1|1, x) - \Pi(1, 0, x)},$$

where Assumption B ensures that $P(t|y, x) \neq 0$ for all $(t, y, x) \in \{0, 1\} \times \{0, 1\} \times X$ in each of the two Bernoulli sampling schemes. It is worth noting that $\Gamma_{d,RR}(x, 0)$ for both $d \in \{CC, CP\}$ is just the covariate-adjusted odds ratio, that is,

$$OR(x) := \frac{\Pi(1|1, x)}{\Pi(0|0, x)} \times \frac{\Pi(0|1, x)}{\Pi(0|0, x)} \times \frac{\Pi(1|1, x)}{\Pi(1|0, x)},$$

which is a popular measure of covariate-adjusted association in case-control studies. Since $OR(x)$ is more descriptive than
The following lemma shows what we could achieve if we had a random sample, that is, if \((Y^*, T^*, X^*)\) were observed.

**Lemma 1 (RR-Benchmark).** If Assumptions A to C are satisfied, then for all \(x \in \mathcal{X}\),

\[
\theta_{RR}(x) = \frac{\mathbb{P}(Y^* = 1 | T^* = 1, X^* = x)}{\mathbb{P}(Y^* = 1 | T^* = 0, X^* = x)}. 
\]

Alternatively, if Assumptions A, B, D, and E are satisfied, then for all \(x \in \mathcal{X}\),

\[
1 \leq \theta_{RR}(x) \leq \frac{\mathbb{P}(Y^* = 1 | T^* = 1, X^* = x)}{\mathbb{P}(Y^* = 1 | T^* = 0, X^* = x)},
\]

where the bounds are sharp.

**Lemma 1** serves two purposes. First, it is useful as a middle step to establish the sharp identifiable bounds under Bernoulli sampling, that is, under Designs 1 (Case-Control) and 2 (Case-Population). Second, it shows benchmark results for the identification of \(\theta_{RR}(x)\) in that it shows the best we can achieve under random sampling through unconfoundedness or monotonicity. Therefore, **Lemma 1** should be compared with Theorems 1–3 that are discussed below.

Point identification under random sampling and strong ignorability is not surprising. Partial identification under random sampling and the monotonicity assumptions is reminiscent of for example Manski and Pepper (2000). However, in our setup, the researcher does not have access to a random sample of \((Y^*, T^*, X^*)\), and therefore, **Lemma 1** is not an identification result. It will serve as a benchmark to show the cost of case-control or case-population studies in terms of identification.

Recall that \(p_0 = \mathbb{P}(Y^* = 1)\) is the true probability of the case, which is an unidentified object under Bernoulli sampling.

**Theorem 1.** Suppose that Assumption A to C are satisfied. Then, for all \(x \in \mathcal{X}\), we have the following.

1. Under case-control sampling, that is, Design 1, we have \(\theta_{RR}(x) = \Gamma_{CC,RR}(x, p_0)\).
2. Under case-population sampling, that is, Design 2, we have \(\theta_{RR}(x) = \text{OR}(x)\).

**Theorem 1** is not identification results in the case of case-control sampling; \(p_0\) is unidentified in Design 1. In contrast, it shows that \(\theta_{RR}(x)\) is point identified under case-population sampling. Therefore, **Design 2** provides an easier environment for causal inference, at least under unconfoundedness. It seems ironic that **Design 2** was referred to as case-control sampling with contamination by Lancaster and Imbens (1996) but that the “contamination” is in fact helpful for identification.

In the case of **Design 1** we do not have point identification, but there is only one simple parameter that is unidentified. Therefore, it is not too difficult to proceed with a partial identification approach. We will further elaborate about this possibility. Before we proceed though, it is worth comparing the case-control case of **Theorem 1** with Holland and Rubin (1988). Specifically, Holland and Rubin (1988) show that under **Design 1**, \(\text{OR}(x)\) is equal to the odds ratio in terms of the potential outcomes if strong ignorability is imposed: that is,

\[
\text{OR}(x) = \frac{\mathbb{P}(Y^*(1) = 1 | X^* = x) \mathbb{P}(Y^*(0) = 0 | X^* = x)}{\mathbb{P}(Y^*(1) = 0 | X^* = x) \mathbb{P}(Y^*(0) = 1 | X^* = x)}.
\]

Equation (5) is an identification result, but its right-hand side expression is not straightforward to interpret. It appears that the reason Holland and Rubin (1988) emphasized the right-hand side expression of (5) instead of the more easily interpretable causal relative risk \(\theta_{RR}(x)\) is that the former is identified by \(\text{OR}(x)\), whereas the latter necessitates addressing the issue that \(p_0\) remains unidentified.

Generally, \(\Gamma_{CC,RR}(x, p_0)\) is different from \(\text{OR}(x)\) = \(\Gamma_{CC,RR}(x, 0)\). However, this issue has been traditionally ignored, because if \(Y^*\) represents a rare event in that \(p_0 \approx 0\), then \(\Gamma_{CC,RR}(x, p_0) \approx \Gamma_{CC,RR}(x, 0)\) by continuity: the assumption of small \(p_0\) is known as the rare disease assumption in epidemiology. However, the quality of the approximation via continuity can quickly decrease as \(p_0\) deviates from zero, that is, the occurrence of \(Y^* = 1\) becomes less uncommon in the population. Therefore, when \(p_0\) is away from zero, a natural alternative approach is to take a partial identification approach, where we target the function \(\Gamma_{CC,RR}(x, \cdot)\) itself, at least within a certain neighborhood of 0.

Below we will write \(f_{A,B}(a, b)\) for the Radon-Nikodym density of \(A, B\) (with respect to some dominating measure). For instance, when \(A\) is discrete and \(B\) is continuous, we will have \(f_{A,B}(a, b) = \mathbb{P}(A = a) f_{B|A}(b|a)\) by using a product of count and Lebesgue measures. Similarly, \(f_{A,B,C}(a, b, c)\) will be used to denote a conditional density of \((A, B)\) at \((a, b)\) given \(C = c\).

**Assumption F.** There is a known value \(\hat{p}\) such that \(p_0 \leq \hat{p}\), where \(\hat{p} \leq 1\) under **Design 1**, and \(\hat{p} \leq \bar{p}\) with

\[
\bar{p}^* := \inf \left\{ \frac{f_{T|X,Y}(t, x|0)}{f_{T|X,Y}(t, x|1)} : t, x \text{ are such that } f_{T|X,Y}(t, x|1) > 0 \right\}
\]

under **Design 2**.

We remark that \(\bar{p}^* \leq 1\): see online appendix F. Under case-control or case-population sampling, \(p_0 = \mathbb{P}(Y^* = 1)\) is generally unidentified, because \(Y^*\) is not randomly observed. Since case-control or case-population sampling is popular when \(Y^* = 1\) is a rare event and therefore a random sample of a modest size tends to contain too few observations of the case of interest, we do not want to rule out the possibility that \(p_0\) is close to zero: it is straightforward though to replace **Assumption F** with the one that \(p_0 \in [\tilde{p}, \bar{p}]\) for some known values of \(\tilde{p}\) and \(\bar{p}\).

If we have an auxiliary sample, from which we learn about \(p_0\), then plugging that piece of information into the case-control sample will resolve the identification problem since \(p_0\) is the only unidentified object here. Even if it is difficult to pin down \(p_0\) exactly, we may have external sources or qualitative information about how prevalent a certain “disease” is, and such information can be used to place an upper bound on \(p_0\). Relying on the researcher’s prior knowledge on an unidentified object has been used in the context of robust estimation as well (e.g., Horowitz and Manski 1995, 1997).

Choosing \(\tilde{p} = 1\) in **Design 1** corresponds to the case where the researcher has no prior information for \(p_0\) at all: we do not...
rule out this possibility. In Design 2, it may be possible to find \( \hat{p} < 1 \) even without having any external source of information at all. To see this point, we note that under Design 2, we must have

\[
\begin{align*}
  f_{T|Y}(t, x|0) &= f_{T,X}(t, x) \\
  &= f_{T,X|Y}(t, x|1)p_0 + f_{T,X|Y}(t, x|0)(1 - p_0) \\
  &= f_{T,Y}(t, x|1)p_0 + f_{T,X|Y}(t, x|0)(1 - p_0),
\end{align*}
\]

where

\[
  f_{T,X|Y}(t, x|0)(1 - p_0) = f_{T,Y}(t, x|0) - f_{T,Y}(t, x|1)p_0 \geq 0
\]

for all \( t, x \). This motivates the definition of \( \hat{p}^* \) in (6).

**Theorem 2.** Suppose that Assumptions A–C, and F are satisfied. Under case-control sampling, that is, Design 1, we have

\[
\min(\Omega(x), \Gamma_{CC,RR}(x, \tilde{p})) \leq \theta_{RR}(x) \leq \max(\Omega(x), \Gamma_{CC,RR}(x, \hat{p})),
\]

and the bounds are sharp.

Theorem 2 is a simple corollary from Theorem 1, where it is addressed that \( p_0 \) is unidentified under case-control sampling. Since \( \Gamma_{CC,RR}(x, p) \) is monotonic in \( p \in [0, \tilde{p}] \), it suffices to consider the two endpoints to obtain sharp bounds, where one of the endpoints is the odds ratio \( \Omega(x) = \Gamma_{CC,RR}(x, 0) \). We also remark that it can be verified that \( \Gamma_{CC,RR}(x, \tilde{p}) \geq 0 \) because \( 0 \leq \rho_{CC}(x, \tilde{p}) \leq 1 \) by definition: this should not be surprising because \( \theta_{RR}(x) \geq 0 \) by definition.

If Assumption D is satisfied in addition, then we can show that \( \Gamma_{CC,RR}(x, \cdot) \) is a decreasing function and therefore it follows that \( \Gamma_{CC,RR}(x, \tilde{p}) \leq \theta_{RR}(x) \leq \Gamma_{CC,RR}(x, 0) = \Omega(x) \) under Design 1. Therefore, the odds ratio represents the maximum causal relative risk that is consistent with what is observed in a case-control study. If there is no information for \( p_0 \) at all, then the lower bound is simply one. Below we will see that the sharp identifiable bounds \( [1, \Omega(x)] \) on \( \theta_{RR}(x) \) can still be obtained without relying on the ignorability assumptions in case-control studies.

Unconfoundedness is a popular assumption for causal inference, but it is not always satisfied in observational studies. Further, unlike the standard case of random sampling, it does not deliver point-identification under case-control studies. Assumptions D and E provide an alternative possibility, where we do not lose much in terms of partial identification.

**Theorem 3.** Suppose that Assumptions A, B, D, and E are satisfied. Then, under both Designs 1 and 2, we have \( 1 \leq \theta_{RR}(x) \leq \Omega(x) \), where the bounds are sharp.

Unlike Theorems 1 and 2, Theorem 3 considers the case where we do not have unconfoundedness but we only impose monotonicity. Now, \( \Omega(x) \) is a sharp upper bound on \( \theta_{RR}(x) \) under both case-control and case-population sampling designs.

It is not explicit in Theorem 3, but its proof shows that the knowledge of \( p_0 \) is potentially useful in Design 1 but not in Design 2. In fact, if \( p_0 = 0 \), then sharp bounds on \( \theta_{RR}(x) \) under Design 1 would be given by \( [1, \Gamma_{CC,RR}(x, p_0)] \), whereas those under Design 2 would still be \( [1, \Omega(x)] \). This difference arises because a few applications of the Bayes rule show that the sharp upper bound under random sampling, that is, the prospective regression ratio \( \mathbb{P}(Y^* = 1|T^* = 1, X^* = x) / \mathbb{P}(Y^* = 1|T^* = 0, X^* = x) \) in Lemma 1, is equal to \( \Gamma_{CC,RR}(x, p_0) \) under Design 1, whereas it is equal to \( \Gamma_{CP,RR}(x, 0) = \Omega(x) \) under Design 2. Therefore, if we do not have a random sample, but we have access only to a case-control sample, then there is an information loss in terms of sharp identifiable bounds on \( \theta_{RR}(x) \). In contrast, a case-population sample is equally informative for \( \theta_{RR}(x) \) as a random sample. Thus, Design 2 provides a better environment for causal inference than Design 1 under monotonicity, similarly to the case of unconfoundedness: see our comments below Theorem 1. The extra challenge in case-control studies can be addressed by the fact that \( \Gamma_{CC,RR}(x, p) \) is decreasing in \( p \). Therefore, the sharp upper bound on \( \theta_{RR}(x) \) under Design 1 is given by the maximum (over \( p \) ) of \( \Gamma_{CC,RR}(x, p) \), which is equal to \( \Gamma_{CC,RR}(x, 0) \) even without using Assumption F.

We now compare Theorem 3 with Theorems 1 and 2. The identification power of strong ignorability depends on the specific sampling design, whereas that of the monotonicity assumptions is independent of which of the two sampling scenarios applies. Specifically, in case-population studies, that is, Design 2, unconfoundedness is informative in that it ensures that \( \theta_{RR}(x) \) is point identified by the odds ratio. However, in case-control studies, that is, Design 1, unconfoundedness only yields interval identification, where the sharp identifiable bounds are the same as what the monotonicity assumptions can deliver if we have no information for \( p_0 \).

### 3.2. Causal Attributable Risk

We now turn to the alternative causal parameter \( \theta_{AR}(x) \). We need some extra notation. For \( (x, p) \in \mathcal{X} \times [0, 1] \) and for \( d \in \{CC, CP\} \), define

\[
\Gamma_{d,AR}(x, p) := \sum_{j=0}^{1} \frac{(-1)^{j+1} \Pi(j|1, x)}{\Pi(j|0, x) + r_d(x, p)(\Pi(j|1, x) - \Pi(j|0, x))},
\]

where \( \Pi(t|y, x) \) and \( r_d(x, p) \) are defined in the beginning of Section 3.1. Note that \( \Gamma_{d,AR}(x, 0) \) is not exactly the odds difference, though it is similar: it is a difference between two ratios of retrospective regressions.

We start with the benchmark case of what if we could observe \( (Y^*, T^*, X^*) \).

**Lemma 2 (AR-Benchmark).** If Assumptions A and C are satisfied, then for all \( x \in \mathcal{X} \),

\[
\theta_{AR}(x) = \mathbb{P}(Y^* = 1|T^* = 1, X^* = x) - \mathbb{P}(Y^* = 1|T^* = 0, X^* = x).
\]

Alternatively, if Assumptions A, B, D, and E are satisfied, then for all \( x \in \mathcal{X} \),

\[
0 \leq \theta_{AR}(x) \leq \mathbb{P}(Y^* = 1|T^* = 1, X^* = x) - \mathbb{P}(Y^* = 1|T^* = 0, X^* = x),
\]

where the bounds are sharp.

Similarly to Lemma 1, Lemma 2 has two purposes. First, it is a middle-step result to establish the sharp identifiable bounds...
on $\theta_{AR}(x)$ when we do not have a random sample but only a sample from either Designs 1 or 2 is available. Second, it shows benchmark results for the identification of $\theta_{AR}(x)$ via unconfoundedness or monotonicity under random sampling. Point identification of $\theta_{AR}(x)$ via strong ignorability under random sampling is now a standard result. If strong ignorability is replaced with the monotonicity assumptions, then the regression difference should be interpreted as a sharp upper bound on the causal attributable risk. Below we extend these results to the cases of case-control and case-population sampling.

**Theorem 4.** Suppose that Assumptions A–C are satisfied. Then, for all $x \in \mathcal{X}$, we have the following.

1. Under case-control sampling, that is, Design 1, we have
   \[
   \theta_{AR}(x) = r_{CC}(x, p_0) \Gamma_{CC,AR}(x, p_0).
   \]
2. Under case-population sampling, that is, Design 2, we have
   \[
   \theta_{AR}(x) = r_{CP}(x, p_0) \Gamma_{CP,AR}(x, 0).
   \]

   Unlike the case of $\theta_{AR}(x)$, $\theta_{AR}(x)$ remains unidentified even in Design 2. This happens because $\Gamma_{CP,AR}(x, 0)$ is a ratio of two terms, where $r_{CP}(x, p_0)$ cancels out, but $\Gamma_{CP,AR}(x, 0)$ is a difference and the common factor $r_{CP}(x, p_0)$ does not disappear. Also, unlike $\theta_{AR}(x)$, the rare disease approximation does not provide anything useful in either of the two sampling schemes: if $p_0 \approx 0$, then $r_{CC}(x, p_0) \approx 0$ and $r_{CP}(x, p_0) \approx 0$ by continuity. However, the partial identification approach still remains useful.

**Theorem 5.** Suppose that Assumptions A–C and F are satisfied. For all $x \in \mathcal{X}$, we have the following.

1. Under case-control sampling, that is, Design 1,
   \[
   \min_{p \in [0, \bar{p}]} r_{CC}(x, p) \Gamma_{CC,AR}(x, p) \leq \theta_{AR}(x) \leq \max_{p \in [0, \bar{p}]} r_{CC}(x, p) \Gamma_{CC,AR}(x, p),
   \]
   where the bounds are sharp.

2. Under case-population sampling, that is, Design 2,
   \[
   \min\{0, r_{CP}(x, \check{p}) \Gamma_{CP,AR}(x, 0)\} \leq \theta_{AR}(x) \leq \max\{0, r_{CP}(x, \check{p}) \Gamma_{CP,AR}(x, 0)\},
   \]
   where the bounds are sharp.

Since $\theta_{AR}(x)$ is a difference of probabilities, it is always between $-1$ and 1. Indeed, we show in the proof that all the bounds in Theorem 5 lie within the interval between $-1$ and 1. Theorem 5 is a simple corollary of Theorem 4: sharpness follows from the fact that $p_0$ is unidentified and that $r_{CC}(x, p) \Gamma_{CC,AR}(x, p)$ and $r_{CP}(x, p) \Gamma_{CP,AR}(x, p)$ are all continuous in $p$. Unlike the case of random sampling, the conditional average treatment effect is only partially identified even under strong ignorability. Also, it is noteworthy that in Design 2, the sign of $\theta_{AR}(x)$ is determined by that of $\Gamma_{CP,AR}(x, 0)$: if we know that $\theta_{AR}(x) \geq 0$, then we know that the conditional average treatment effect is at most $r_{CP}(x, \check{p}) \Gamma_{CP,AR}(x, 0)$.

We now consider replacing unconfoundedness with the monotonicity assumptions.

**Theorem 6.** Suppose that Assumptions A, B, and D to F are satisfied. Then, for all $x \in \mathcal{X}$, we have the following.

1. Under case-control sampling, that is, Design 1,
   \[
   0 \leq \theta_{AR}(x) \leq \max_{p \in [0, \bar{p}]} r_{CC}(x, p) \Gamma_{CC,AR}(x, p),
   \]
   where the bounds are sharp.

2. Under case-population sampling, that is, Design 2,
   \[
   0 \leq \theta_{AR}(x) \leq r_{CP}(x, \check{p}) \Gamma_{CP,AR}(x, 0),
   \]
   the bounds are sharp.

Similarly to our comments below Theorem 3, knowledge of $p_0$ is potentially useful to improve the bounds given in Theorem 6: this point will be relevant when we discuss aggregation in the following section. This is so because, by the Bayes rule, the difference between the two prospective regression functions that appear in Lemma 2 can be shown to be equal to $r_{CC}(x, p_0) \Gamma_{CC,AR}(x, p_0)$ under Design 1 and to $r_{CP}(x, p_0) \Gamma_{CP,AR}(x, 0)$ under Design 2, respectively. However, $p_0$ is unrestricted in general, and hence maximizing over $p_0 \in [0, \bar{p}]$ under Assumption F delivers the sharp upper bounds.

The bounds in Theorem 6 are comparable with those in Theorem 5. In case-control or case-population sampling, strong ignorability is not as powerful as in random sampling. First, strong ignorability does not deliver point identification of the conditional average treatment effect. Second, the monotonicity assumptions do restrict the sign of $\theta_{AR}(x)$, but, otherwise, they have the same amount of information as the strong ignorability assumptions in terms of the maximum admissible value of $\theta_{AR}(x)$.

### 4. Aggregation

Conditioning on a specific value of the covariate vector and aiming at $\theta_{RR}(x)$ or $\theta_{AR}(x)$ as in Theorems 2, 3, 5, and 6 is one natural approach to deal with potential heterogeneity in the causal treatment effect. However, the corresponding bounds as functions of $x$ (e.g., OR($x$)) are complicated objects, and they are difficult to estimate with high precision when $X^*$ is multi-dimensional.

To avoid the curse of dimensionality, it is popular in case-control studies to adopt logistic regression. Some authors have alternatively parameterized the odds ratio function itself in case-control studies, focusing on establishing a doubly robust estimator of the odds ratio: see for example, Chen (2007) and Tchetgen Tchetgen (2013). Direct parameterization of $\Gamma_{d,AR}(x, p)$ appears to be uncommon though.

Parametric assumptions are convenient, but they are restrictive: for example, OR($x$) is generally an unknown function of $x$ that can be highly nonlinear. Instead of introducing any parameterization, aggregation over the population distribution of the covariates can be a useful approach to obtain a robust summary measure.

If one wants to report an aggregated parameter such as $\int_X \theta_{AR}(x) \omega(x) dx$ for some weight function $\omega$, sharp bounds can be obtained by taking max/min over $p_0$ after aggregation. The most natural choice of the weight function $\omega$ is probably the true population density of $X^*$. The distribution of $X^*$ is unidentified in case-control studies, but the situation is not too bad because the only unidentified object is, again, $p_0$. 

**Theorem 6.** Suppose that Assumptions A, B, and D to F are satisfied. Then, for all $x \in \mathcal{X}$, we have the following.
Consider the following aggregated parameters:
\[ \tilde{\theta}_{\text{RR}} := \int_{\mathcal{X}} \log \{ \theta_{\text{RR}}(x) \} f_X(x) \, dx \quad \text{and} \]
\[ \tilde{\theta}_{\text{AR}} := \int_{\mathcal{X}} \theta_{\text{AR}}(x) f_X(x) \, dx. \quad (7) \]

\( \tilde{\theta}_{\text{AR}} \) is the standard average treatment effect. For \( \tilde{\theta}_{\text{RR}} \), we use the logarithm of \( \theta_{\text{RR}}(X^*) \) to take an average. Since \( \mathbb{E} [\log \text{OR}(X^*)] \leq \log \mathbb{E} [\text{OR}(X^*)] \) by Jensen's inequality, the average of the logarithm is less likely to be affected unduly by outliers. We also note that it is more conventional to work with the logarithm of the odds ratio than the odds ratio itself. If one still prefers aggregating \( \theta_{\text{RR}}(x) \) itself, it is straightforward to modify our methodology by using the same principle outlined in this section.

Our approach is to use the fact that the only missing piece in case-control or case-population samples is \( P_0 \). We first derive sharp identifiable bounds on \( \theta_{\text{RR}}(x) \) and \( \theta_{\text{AR}}(x) \) with \( P_0 \) given. We then aggregate over the distribution of \( X^* \), which depends on \( P_0 \) in case-control studies. Specifically, we use the fact that for all \( x \in \mathcal{X} \),
\[ f_X(x) = \begin{cases} f_{X \mid Y}(x|1)P_0 + f_{X \mid Y}(x|0)(1 - P_0) & \text{in case-control studies}, \\ f_{X \mid Y}(x|0) & \text{in case-population studies}. \end{cases} \]

We can then rely on Assumption F to address the fact that \( P_0 \) is unidentified. For this purpose, we can maximize or minimize over \( P_0 \in [0, \bar{p}] \) to obtain bounds, or, more informatively, we can plot the whole bound functions on \( [0, \bar{p}] \); choosing the maximal value that is allowed for \( \bar{p} \) (e.g., \( \bar{p} = 1 \) in case-control studies) corresponds to the case where we have no information for \( P_0 \). This line of reasoning leads to the main results in this section.

We will use the following objects: for \( d \in \{ \text{CC}, \text{CP} \} \),
\[ \Psi_{d, \text{RR}}(p, y) := \mathbb{E} [\log \Gamma_{d, \text{RR}}(X, p) | Y = y]. \]

The logarithm in the definition of \( \Psi_{d, \text{RR}}(p, y) \) is because \( \tilde{\theta}_{\text{RR}} \) is the aggregation of \( \log \theta_{\text{RR}}(x) \). If one wants to bound \( \int_{\mathcal{X}} \theta_{\text{RR}}(x) f_X(x) \, dx \), then changing the definition of \( \Psi_{d, \text{RR}}(p, y) \) to \( \mathbb{E} [\Gamma_{d, \text{RR}}(X, p) | Y = y] \) will do. Also, we note that \( \int_{\mathcal{X}} \theta_{\text{RR}}(x) f_X(x) \, dx \) differs from the ratio of unconditional counterfactual probabilities. Let
\[ \Psi_{\text{CC,AR}}(p, y) := \mathbb{E} [\text{OR}(X, p) \Gamma_{\text{CC,AR}}(X, p) | Y = y], \]
\[ \Psi_{\text{CP,AR}}(p, y) := \mathbb{E} [\text{OR}(X, p) \Gamma_{\text{CP,AR}}(X, p) | Y = 0], \]
where we note that \( \Psi_{\text{CP,AR}}(p) \) is a simple linear function of \( p \) by definition. Finally, for \( k \in \{ \text{RR}, \text{AR} \} \), define \( C_{\text{CC,k}}(p) \) by a convex combination of \( \Psi_{\text{CC,k}}(p, 1) \) and \( \Psi_{\text{CC,k}}(p, 0) \); that is, \( C_{\text{CC,k}}(p) := \Psi_{\text{CC,k}}(p, 1)p + \Psi_{\text{CC,k}}(p, 0)(1 - p) \).

**Theorem 7.** Suppose that Assumptions A–C, and F are satisfied. We then have the following.

1. Under case-control sampling, that is, Design 1, the sharp identified bounds on \( \tilde{\theta}_{\text{RR}} \) and \( \tilde{\theta}_{\text{AR}} \) are given by
\[ \min_{p \in [0, \bar{p}]} C_{\text{CC,RR}}(p) \leq \tilde{\theta}_{\text{RR}} \leq \max_{p \in [0, \bar{p}]} C_{\text{CC,RR}}(p), \]
\[ \min_{p \in [0, \bar{p}]} C_{\text{CC,AR}}(p) \leq \tilde{\theta}_{\text{AR}} \leq \max_{p \in [0, \bar{p}]} C_{\text{CC,AR}}(p). \]

2. Under case-population sampling, that is, Design 2, we have \( \tilde{\theta}_{\text{RR}} = \Psi_{\text{CC,RR}}(0, 0) \), where we remark that this point identification result does not require Assumption F. Further, the sharp identified bounds on \( \tilde{\theta}_{\text{AR}} \) are given by
\[ \min \{ 0, \Psi_{\text{CP,AR}}(\bar{p}) \} \leq \tilde{\theta}_{\text{AR}} \leq \max \{ 0, \Psi_{\text{CP,AR}}(\bar{p}) \}. \]

**Theorem 8.** Suppose that Assumptions A, B, and D to F are satisfied. Then, we have the following.

1. Under the case-control sampling, that is, Design 1, the sharp identified bounds on \( \tilde{\theta}_{\text{RR}} \) and \( \tilde{\theta}_{\text{AR}} \) are given by
\[ 0 \leq \tilde{\theta}_{\text{RR}} \leq \max_{p \in [0, \bar{p}]} C_{\text{CC,RR}}(p) \quad \text{and} \]
\[ 0 \leq \tilde{\theta}_{\text{AR}} \leq \max_{p \in [0, \bar{p}]} C_{\text{CC,AR}}(p). \]

2. Under the case-population sampling, that is, Design 2, the sharp identified bounds on \( \tilde{\theta}_{\text{RR}} \) and \( \tilde{\theta}_{\text{AR}} \) are given by
\[ 0 \leq \tilde{\theta}_{\text{RR}} \leq \Psi_{\text{CP,RR}}(0, 0) \quad \text{and} \quad 0 \leq \tilde{\theta}_{\text{AR}} \leq \Psi_{\text{CP,AR}}(\bar{p}), \]
where we remark that the bounds on \( \tilde{\theta}_{\text{RR}} \) do not rely on Assumption F.

Generally, in both cases of strong ignorability and monotonicity, case-population sampling provides an easier environment for causal inference than case-control studies: \( \Psi_{\text{CC,RR}}(0, 0) \) does not depend on \( p \) and \( \Psi_{\text{CP,AR}}(p) \) is linear in \( p \). Also, the bounds under strong ignorability are all comparable with those under monotonicity: the upper bounds have the same form under strong ignorability as under monotonicity except that the monotonicity assumptions impose restrictions on the direction of the causal effect.

Theorems 7 and 8 show that \( \tilde{\theta}_{\text{RR}} \) suites better case-control or case-population studies than \( \tilde{\theta}_{\text{AR}} \), especially when the case is potentially rare, despite the popularity of the latter in random sampling. Specifically, \( \text{OR}_{\text{CC}}(x, p) \) and \( \text{OR}_{\text{CP}}(x, p) \) should be taken into account for \( \tilde{\theta}_{\text{AR}} \), but they are irrelevant for \( \tilde{\theta}_{\text{RR}} \). This is an important difference because \( \text{OR}_{\text{CC}}(x, 0) = \text{OR}_{\text{CP}}(x, 0) = 0 \), which implies that the bounds on \( \tilde{\theta}_{\text{AR}} \) cannot be tighter under strong ignorability than under monotonicity. In order to see the point more clearly, consider the case of case-control studies, that is, Design 1, and suppose that \( \max_{p \in [0, \bar{p}]} \{ \Psi_{\text{CC,AR}}(p, 1)p + \Psi_{\text{CC,AR}}(p, 0)(1 - p) \} > 0 \) so that the upper bound on \( \tilde{\theta}_{\text{AR}} \) is positive both under strong ignorability and under monotonicity. In this case, the lower bound on \( \tilde{\theta}_{\text{AR}} \) under strong ignorability can never be strictly positive because \( \Psi_{\text{CC,AR}}(0, y) \) is trivially equal to zero. In other words, strong ignorability does provide a more informative environment than monotonicity but only in the sense that the former does not restrict the sign of \( \tilde{\theta}_{\text{AR}} \). Once the sign of \( \tilde{\theta}_{\text{AR}} \) is given, then there is nothing extra the strong ignorability assumptions offer relative to the monotonicity setup in understanding the average treatment effect. The same is true for the case-population case, that is, Design 2.

If we focus on \( \tilde{\theta}_{\text{RR}} \), then the average of the log odds ratios, that is, \( \beta(y) := \Psi_{\text{CC,RR}}(0, y) - \Psi_{\text{CC,RR}}(0, \bar{y}) = \mathbb{E} [\log \text{OR}(X) | Y = y] \) becomes the central object for estimation and inference. For instance, in Design 2, all we need is \( \beta(0) \), which can be interpreted as \( \tilde{\theta}_{\text{RR}} \) itself or its sharp upper bound, depending on whether we assume strong ignorability or monotonicity,
respectively. In Design 1, if Assumption D is imposed, then \( \Psi_{CC,RR}(p, y) \) can be shown to be decreasing in \( p \), and therefore we have \( C_{CC,RR}(p) \leq \beta(1)p + \beta(0)(1 - p) \). Since the right-hand side is linear in \( p \), we can easily conduct inference on \( \delta_{RR} \) uniformly in \( p \in [0, \bar{p}] \) by using \( \beta(y) \), though this can be conservative.

The log odds ratio \( \log \text{OR}(x) \) has been a popular measure of association in case-control studies, and \( \beta(y) \) is an aggregation of it by using the identified distribution of \( X \) given \( Y = y \). Jun and Lee (2023) establish the semiparametric efficiency bound for \( \mathbb{E}(\log \text{OR}(X)) \) and suggest efficient estimators that accommodate high-dimensional machine learning estimators in the first stage. For low-dimensional \( X \), a straightforward algorithm for efficient estimation of \( \beta(y) \) is available, and it can be easily implemented using standard software. The algorithm is described in online appendix B2.

5. Causal Inference Under Monotonicity

In this section, we discuss how to carry out causal inference on the aggregated parameters \( \delta_{RR} \) and \( \delta_{AR} \) under the MTR and MTS assumptions: inference under strong ignorability can be done by the same principles. In our discussion below, \( z(1 - \alpha) \) will be the \( 1 - \alpha \) quantile of the standard normal distribution.

We first consider relative risk, for which we use \( \exp(\delta_{RR}) \) as the parameter of interest: see our discussion right below equation (7). Our basis for inference is Theorem 8. Let \( \beta(y) := \Psi_{CP,RR}(0, y) = \mathbb{E}(\log \text{OR}(X)|Y = y) \) for \( y = 0, 1 \).

Inference is easier when we have a case-population sample: all we need is \( \beta(0) \). Since we have \( 1 \leq \exp(\delta_{RR}) \leq \exp(\beta(0)) \) by Theorem 8, a \( 1 - \alpha \) confidence interval for \( \exp(\delta_{RR}) \) can be constructed by \([1, \exp(\hat{\delta}(0) + z(1 - \alpha)\hat{s}(0))]\), where \( \hat{\delta}(0) \) is an asymptotically normal estimator of \( \beta(0) \), and \( \hat{s}(0) \) is its standard error.

In the case of case-control sampling, that is, Design 1, we should base our inference on \( C_{CC,RR}(\cdot) \). However, \( C_{CC,RR}(p) \) is nonlinear in \( p \), and hence it is difficult to obtain a confidence band uniformly in \( p \in [0, \bar{p}] \). We propose two solutions. One is just to use one-sided pointwise confidence bands using bootstrap, akin to algorithm 2 in online appendix D, where we focus on pointwise inference for \( \delta_{AR} \). The other is to take a conservative approach by using the fact that \( C_{CC,RR}(p) \leq \tilde{\beta}(p) := \beta(1)p + \beta(0)(1 - p) \). Specifically, in online appendix D, we show that

\[
\mathbb{P}[\forall p \in [0, 1], \exp(\tilde{\beta}(p)) \leq \exp(\hat{\beta}(1) + (1 - p)\hat{\beta}(0) + u(1 - \alpha)) \geq 1 - \alpha,
\]

where \( u(1 - \alpha) := z(1 - \alpha/2) \max[\hat{s}(0), \hat{s}(1)] \) with \( \hat{s}(y) \) is the standard error of \( \hat{\beta}(y) \), the asymptotically normal estimator of \( \beta(y) \).

We now turn to inference on \( \delta_{AR} \). The case-population sample provides an easier environment again: we can exploit the fact that \( \Psi_{CP,AR}(p) \) is a simple linear function with the form of \( \Psi_{CP,AR}(p) := \rho_{CP} \), where \( \rho_{CP} \) is implicitly defined here and does not depend on \( p \). For more details, see online appendix D.

Inference on \( \delta_{AR} \) with a case-control sample relies on the function \( C_{CC,AR}(\cdot) \), and its nonlinearity in \( p \) makes it difficult to construct a uniform confidence band. Since \( \Psi_{CC,AR}(\cdot, y) \) is not monotonic, the conservative approach we discussed for \( \delta_{RR} \) does not apply here. Therefore, we propose using one-sided pointwise confidence intervals, for which we use Efron’s bias-corrected percentile intervals. Computational details for implementation are given in online appendix D.

6. Empirical Examples

6.1. Case-Control Sampling: Entering a Very Selective University

We consider quantifying the causal effect of attending private school on entering a very selective university by using the Pakistan data collected by Delavande and Zafar (2019). This is survey data from male students who were already enrolled in different types of universities in Pakistan, all located in Islamabad/Rawalpindi and Lahore. Delavande and Zafar (2019) include two Western-style universities, one Islamic university, and four madrassas, but we focus on the two Western-style ones in our analysis: between the two universities, Delavande and Zafar (2019) call the more expensive, selective, and reputable university “Very Selective University” (VSU) and the other simply “Selective University” (SU). Therefore, we restrict the population of interest to those who entered either VSU or SU, and we define the binary outcome to be whether a student entered VSU. The binary treatment we consider is whether a student attended private school before university. Since the students in the sample were already enrolled in either VSU or SU at the time of the survey, we have a case-control sample, that is, our Design 1.

Table 1 shows the likelihood of entering VSU by private school attendance before university. The empirical odds ratio is 1.38.

In this example, the unconfoundedness assumption is unlikely to hold, because those who attended private school before university are likely to have more resourceful parents. This concern may not completely disappear even if we control for parental income and wealth because of the presence of unobserved parental abilities and resources that could affect their children’s university choice. However, the MTR and MTS assumptions are still plausible: private school is probably no inferior input to university preparations (hence, MTR), and those who actually chose to attend private school probably care about their future college choice no less than those who did not (hence, MTS). Then, the odds ratio of 1.38 can be interpreted as a sharp upper bound on causal relative risk; therefore, the effect of attending private school seems, at best, modest.

Now, we consider controlling for family background variables. Specifically, we include an indicator for at least one college-educated parent and parents’ monthly income as covariates. Table 2 reports estimation results for the aggregated log odds ratio within each of the case and the control: both \( \beta(y) \) and
exp[\beta(y)] convey the same information, but exp[\beta(y)] is easier to interpret because it is comparable to the usual odds ratio in terms of its scale. The fact that  \hat{\beta}(1) and  \hat{\beta}(0) are notably different suggests that the amount of heterogeneity among individuals may be substantial. The confidence intervals are computed based on the MTR and MTS: hence, they are one-sided.

We now consider the methods described in Section 5, that is, causal inference on the aggregated relative and attributable risk (RR and AR, respectively) in terms of the population distribution of the covariates. We rely on the MTR and MTS assumptions to interpret our results as upper bounds. For AR, we use the same covariates as in RR. The number of the bootstrap replications was 10,000.

Figure 1 summarizes the results: the left (right) panel shows RR (AR). The case probability \( p_0 \) of entering VSU in the population is not identified in this dataset. But we can trace out the upper bounds as the value of \( p_0 \) varies between 0 and 1.

Consider the left panel of Figure 1, that is, RR, where we take the conservative approach and plot \( \hat{\beta}(p) = \beta(1)p + \beta(0)(1-p) \). If we take the point estimate at face value, attending private school increases the chance of entering VSU by a factor of at most 1.26. Even in terms of the confidence intervals, it seems highly unlikely that the impact is more than a factor of 2. The right panel of Figure 1 shows AR. The graph shows an inverted U-shape, because \( r_{CC}(x,p)I_{CC,AR}(x,p) = 0 \) whenever \( p \) is either 0 or 1. The maximum point estimate of the upper bound is 0.044, while the maximum value of the confidence intervals is 0.153. Therefore, it seems highly unlikely that attending private school increases the chance of entering VSU by more than 16 percent.

None of our results require strong ignorability or the rare-disease assumption. Our conclusion of a relatively small positive effect of attending private school on entering VSU, if it exists at all, is reminiscent of existing results in labor economics.
that find access to private schools have only modest effects on children’s performance (see, e.g., Epple, Romano, and Urquiola 2017; MacLeod and Urquiola 2019).

6.2. Case-Population Sampling: Joining a Criminal Gang

We revisit Carvalho and Soares (2016), who combine the 2000 Brazilian Census with a unique survey of drug-trafficking gangs in favelas (slums) of Rio de Janeiro; therefore, their dataset is an example of case-population sampling, that is, our Design 2. In their study, they use the method of Lancaster and Imbens (1996) to estimate a model of selection into the gang by using race, age, illiteracy, house ownership, and religiosity. They note that the five characteristics are likely to be predetermined while years of schooling may be endogenous to entry, that is, joining the gang may lead members to drop out of school. Indeed, 90% of gang members are not in school, whereas 46% of men aged 10–25 are not in school. They find that “younger individuals, from lower socioeconomic background (black, illiterate, and from poorer families) and with no religious affiliation are more likely to join drug-trafficking gangs.”

Table 3 provides summary statistics of the sample. We regard currently not attending school as the treatment variable of interest. Unconfoundedness is not plausible because of the endogeneity of schooling that we mentioned earlier. Furthermore, it is plausible that unmeasured factors such as family support could affect both treatment and outcome. However, not being in school may increase the chance of exposure to gang-related activities, and those who chose to be in school may be the ones who care for consequences no less than those who chose not to be; therefore, the MTR and MTS follow, respectively.

Table 4 presents estimation results for $\beta(y)$ and $\exp(\beta(y))$, for which we control for the same covariates as Carvalho and Soares (2016). Unlike Table 2, $y = 0$ now corresponds to the entire population. Therefore, $\beta(0)$ itself is the log odds ratio aggregated over the population, which is the sharp upper bound on the aggregation of the log causal relative risk, that is, $\bar{\theta}_{RR}$. The point estimate of $\beta(0)$ is 2.71, and that of $\exp(\beta(0))$ is 15.01, which suggests that the chance of those who are not in school joining a gang may be (up to) 15 times as large as that of those who are.

Our discussion above can be supplemented by checking the causal AR. At the three points of $p_0 \in \{0.05, 0.10, 0.15\}$ that Carvalho and Soares (2016) considered, the point estimates and the end-points of the uniform confidence interval (in parentheses) for the upper bound on the causal AR are 0.33 (0.43), 0.66 (0.86), and 0.99 (1), respectively. The uniform confidence band is based on 1000 bootstrap replications. Note that the confidence band is truncated at one, because AR cannot be larger than one.

Overall, our results are suggestive of potentially large impacts of keeping young men in school in order to discourage them to participate in criminal activities. Further research based on careful study designs would be necessary to reach a more definitive answer.

6.3. Random Sampling: Physician’s Hours

Fang and Gong (2017) construct estimates for physicians’ hours spent on Medicare beneficiaries and find that about 3% of physicians billed for more than 100 hr per week. They refer to these physicians as flagged physicians. Fang and Gong (2017, p. 573) state that “flagged physicians are slightly more likely to be male, non-MD, more experienced, and provide fewer E/M services. Importantly, they work in substantially smaller group practices (if at all), and have fewer hospital affiliations.” We use their study to illustrate the findings in this article. Specifically, the outcome variable is whether a physician billed for more than 100 hr per week in either 2012 or 2013, the treatment variable is a binary indicator showing whether or not the number of group practice numbers is less than 6, which is the median size in the data, and the covariates include an indicator for male, an indicator for doctor of medicine (MD), and experience in years (cubic polynomial).

The original dataset in Fang and Gong (2017) is updated in Fang and Gong (2020) after Matsumoto (2020) pointed out data and coding errors in the original work. In our analysis, we use the updated dataset. Table 5 summarizes the sample, which consists of 78,165 physicians who billed at least 20 hr per week.

Treating this dataset as a random sample, we extract a case-control dataset: the case sample is composed of 2261 flagged physicians; the control sample of equal size is randomly drawn.
without replacement from the pool of physicians who were never flagged. Analogously, a case-population dataset is obtained by combining the case sample with a population sample of equal size that is randomly drawn without replacement from all observations and its flagged status is coded missing.

It is highly unlikely that the group practice size is exogenous conditional on a small number of covariates; hence, we rely on the monotonicity assumptions (i.e., MTR and MTTS) and focus on the upper bounds on the relative and attributable risks.

Table 6 reports the estimates of $\exp(\hat{\beta}(y))$ and their one-sided confidence intervals of $\exp(\hat{\beta}(y))$ for each sampling scheme. The lower bound is 1 because of the MTR assumption and averaging is done for a relevant population in each column. Because the proportion of the flagged physicians is less than 0.03, we invoke the rare disease assumption and regard $\exp(\hat{\beta}(y))$ as an approximation to the upper bound on RR. Although there are some noticeable differences across different columns, the estimates are similar. This is consistent with the identification result that the price to pay is less for identification of RR when we move from random sampling to outcome-dependent sampling. The story is different if we focus on AR. Table 7 shows that the upper bounds on AR under outcome-dependent sampling are much larger than those under random sampling, especially when $\hat{p} = 0.1$.

### Supplementary Materials

In appendices A to C, we discuss aggregation of $\theta_{AR}(\cdot)$ without taking the logarithm, efficient estimation of $\hat{\beta}(y)$ for $y = 0, 1$, and the results of a small Monte Carlo experiment, respectively. Appendix D gives details on inference issues omitted in section 5, and appendix E presents proofs.

### Acknowledgments

The authors would like to thank the editor, an associate editor, two anonymous referees, Guido Imbens, Chuck Manski, Francesca Molinari and seminar participants at Cemmap, Oxford, and Penn State for helpful comments and Leandro Carvalho and Rodrigo Soares for sharing their dataset and help.

### Disclosure Statement

No potential conflict of interest was reported by the author(s).

### Funding

This work was supported in part by the European Research Council (ERC-2014-COG-646917-ROMIA) and by the UK Economic and Social Research Council (ESRC) through research grant (ES/P008909/1) to the CeMMAP.

### ORCID

Sokbae Lee [http://orcid.org/0000-0003-4080-7733]

### References

Bhattacharya, J., Shaikh, A. M., and Vytlaclil, E. (2008), “Treatment Effect Bounds Under Monotonicity Assumptions: An Application to Swan-Ganz Catheterization,” *American Economic Review: Papers and Proceedings*, 98, 351–56. [998]

Breslow, N. E. (1996), “Statistics in Epidemiology: The Case-Control Study,” *Journal of the American Statistical Association*, 91, 14–28. [999]

Breslow, N. E., and Day, N. E. (1980), *Statistical Methods in Cancer Research I. The Analysis of Case-Control Studies* (Vol. 1), Lyon, France: International Agency for Research on Cancer. [998]

Breslow, N. E., Robins, J. M., and Wellner, J. A. (2000), “On the Semi-Parametric Efficiency of Logistic Regression Under Case-Control Sampling,” *Bernoulli*, 6, 447–455. [999]

Carvalho, L. S., and Soares, R. R. (2016), “Living on the Edge: Youth Entry, Career and Exit in Drug Selling Gangs,” *Journal of Economic Behavior & Organization*, 121, 77–98. [998,1000,1007]

Chen, H. (2007), “A Semiparametric Odds Ratio Model for Measuring Association,” *Biometrics*, 63, 413–421. [1003]

Chen, K. (2001), “Parametric Models for Response-Biased Sampling,” *Journal of the Royal Statistical Society*, Series B, 63, 775–789. [999]

Choi, D. (2017), “Estimation of Monotone Treatment Effects in Network Experiments,” *Journal of the American Statistical Association*, 112, 1147–1155. [998]

Currie, J., and Neidell, M. (2005), “Air Pollution and Infant Health: What Can We Learn from California’s Recent Experience?” *Quarterly Journal of Economics*, 120, 1003–1030. [998]

Delavande, A., and Zafar, B. (2019), “University Choice: The Role of Expected Earnings, Nonpecuniary Outcomes, and Financial Constraints,” *Journal of Political Economy*, 127, 2343–2393. [998,1005]

Didelez, V., and Evans, R. J. (2018), “Chapter 6 Causal Inference from Case-Control Studies,” in *Handbook of Statistical Methods for Case-Control Studies*, eds. Ø. Borgan, N. Breslow, N. Chatterjee, M. H. Gail, A. Scott, and C. J. Wild, pp. 87–115, Boca Raton, FL: CRC Press. [999]

Domowitz, I., and Sartain, R. L. (1999), “Determinants of the Consumer Bankruptcy Decision,” *Journal of Finance*, 54, 403–420. [998]

Epplle, D., Romano, R. E., and Urquidi, M. (2017), “School Vouchers: A Survey of the Economics Literature,” *Journal of Economic Literature*, 55, 441–92. [1007]

Fang, H., and Gong, Q. (2017), “Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked,” *American Economic Review*, 107, 562–91. [1007]

Fang, H., and Gong, Q. (2020), “Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked: Reply,” *American Economic Review*, 110, 4004–4010. [1007]

Gabriel, E. E., Sachs, M. C., and Sjölander, A. (2022), “Causal Bounds for Outcome-Dependent Sampling in Observational Studies,” *Journal of the American Statistical Association*, 117, 939–950. [999]

Holland, P. W., and Rubin, D. B. (1988), “Causal Inference in Retrospective Studies,” *Evaluation Review*, 12, 203–231. [998,1001]

Horowitz, J. L., and Manski, C. F. (1995), “Identification and Robustness with Contaminated and Corrupted Data,” *Econometrica*, 63, 281–302. [1001]

Horowitz, J.-W., and van de Meent, pp. 5980–5996, PMLR. [1005]

Jiang, Z., Chiba, Y., and VanderWeele, T. J. (2014), “Monotone Confounding, Monotone Treatment Selection and Monotone Treatment Response,” *Journal of Causal Inference*, 2, 1–12. [998]

Jun, S. J., and Lee, S. (2023), “Average Adjusted Association: Efficient Estimation with High Dimensional Confounders,” in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, Volume 206 of Proceedings of Machine Learning Research, eds. F. Ruiz, J. Dy, and J.-W. van de Meent, pp. 5980–5996, PMLR. [1005]

Kim, W., Kwon, K., Kwon, S., and Lee, S. (2018), “The Identification Power of Smoothness Assumptions in Models with Counterfactual Outcomes,” *Quantitative Economics*, 9, 617–642. [998]

Kreider, B., Pepper, J. V., Gundersen, C., and Jolliffe, D. (2012), “Identifying the Effects of SNAP (Food Stamps) on Child Health Outcomes When Participation is Endogenous and Misreported,” *Journal of the American Statistical Association*, 107, 958–975. [998]

Kuroki, M., Cai, Z., and Geng, Z. (2010), “Sharp Bounds on Causal Effects in Case-Control and Cohort Studies,” *Biometrika*, 97, 123–132. [999]

Lancaster, T., and Imbens, G. (1996), “Case-Control Studies with Contaminated Controls,” *Journal of Econometrics*, 71, 145–160. [999,1001,1007]

Machado, C., Shaikh, A. M., and Vytlaclil, E. J. (2019), “Instrumental Variables and the Sign of the Average Treatment Effect,” *Journal of Econometrics*, 212, 522–555. [998]
MacLeod, W. B., and Urquiola, M. (2019), "Is Education Consumption or Investment? Implications for School Competition," *Annual Review of Economics*, 11, 563–589. [1007]
Manski, C. F. (1995), *Identification Problems in the Social Sciences*, Cambridge, MA: Harvard University Press. [998]
——— (1997), "Monotone Treatment Response," *Econometrica*, 65, 1311–1334. [998,1000]
——— (2003), *Partial Identification of Probability Distributions*, New York: Springer-Verlag. [998]
——— (2007), *Identification for Prediction and Decision*, Cambridge, MA: Harvard University Press. [998]
Manski, C. F., and Lerman, S. R. (1977), "The Estimation of Choice Probabilities from Choice based Samples," *Econometrica*, 45, 1977–1988. [1000]
Manski, C. F., and Pepper, J. V. (2000), "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010. [998,1000,1001]
Månsson, R., Joffe, M. M., Sun, W., and Hennessy, S. (2007), "On the Estimation and Use of Propensity Scores in Case-Control and Case-Cohort Studies," *American Journal of Epidemiology*, 166, 332–339. [998]
Matsumoto, B. (2020), "Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked: Comment," *American Economic Review*, 110, 3991–4003. [1007]
Molinari, F. (2020), "Microeconometrics with Partial Identification," in *Handbook of Econometrics, Volume 7A* Vol. 7 of Handbook of Econometrics, eds. S. N. Durlauf, L. P. Hansen, J. J. Heckman and R. L. Matzkin, pp. 355–486, Amsterdam: Elsevier. [998]
Okumura, T., and Usui, E. (2014), "Concave-Monotone Treatment Response and Monotone Treatment Selection: With an Application to the Returns to Schooling," *Quantitative Economics*, 5, 175–194. [998]
Pearl, J. (2009), *Causality*, Cambridge: Cambridge University Press. [998]
Rose, S. (2011), "Causal Inference for Case-Control Studies," PhD thesis, Biostatistics of the UC-Berkeley. [998]
Rosenfeld, B. (2017), "Reevaluating the Middle-Class Protest Paradigm: A Case-Control Study of Democratic Protest Coalitions in Russia," *American Political Science Review*, 111, 637–652. [1000]
Tamer, E. (2010), "Partial Identification in Econometrics," *Annual Review of Economics*, 2, 167–195. [998]
Tchetgen Tchetgen, E. J. (2013), "On a Closed-Form Doubly Robust Estimator of the Adjusted Odds Ratio for a Binary Exposure," *American Journal of Epidemiology*, 177, 1314–1316. [1003]
Van der Laan, M. J., and Rose, S. (2011), *Targeted Learning: Causal Inference for Observational and Experimental Data*, New York: Springer. [998]
VanderWeele, T. J., and Robins, J. M. (2009), "Properties of Monotonic Effects on Directed Acyclic Graphs," *Journal of Machine Learning Research*, 10, 699–718. [998]
Vylacil, E., and Yildiz, N. (2007), "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75, 757–779. [998]
Xie, J., Lin, Y., Yan, X., and Tang, N. (2020), "Category-Adaptive Variable Screening for Ultra-High Dimensional Heterogeneous Categorical Data," *Journal of the American Statistical Association*, 115, 747–760. [999]