CAUSAL INFERENCE IN CASE-CONTROL STUDIES*

SUNG JAE JUN† AND SOKBAE LEE‡

PENN STATE UNIV. COLUMBIA UNIV. AND IFS

October 28, 2020

Abstract: We investigate partial identification of causal relative and attributable risk—the ratio of two counterfactual proportions and the difference between them—in case-control and case-population studies. The odds ratio is shown to be a sharp upper bound on causal relative risk under the monotone treatment response and monotone treatment selection assumptions, without resorting to strong ignorability, nor to the rare-disease assumption. Sharp bounds on causal attributable risk are also obtained under the same assumptions. Paying special attention to the (conditional) odds ratio, we propose a semiparametrically efficient estimator of the aggregated (log) odds ratio. Further, we develop easy-to-implement causal inference procedures for relative and attributable risk. Finally, we showcase our methodology by applying it to two unique datasets in the literature. We find that attending private school may have little effect on entering a very selective university in Pakistan and that dropping out of school could substantially increase relative and attributable risk of joining a criminal gang in Brazil.

Key Words: relative risk, attributable risk, causality, monotonicity, outcome-based sampling, partial identification

JEL Classification Codes: C21, C55, C83

---

*We would like to thank Guido Imbens, Chuck Manski and seminar participants at Cemmap, Oxford, and Penn State for helpful comments and Leandro Carvalho and Rodrigo Soares for sharing their dataset and help. 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.

†Department of Economics, Penn State University, 619 Kern Graduate Building, University Park, PA 16802, suj14@psu.edu

‡Department of Economics, Columbia University, 1022 International Affairs Building, 420 West 118th Street, New York, NY 10027, sl3841@columbia.edu

1. INTRODUCTION

Empirical researchers often find it useful to work with outcome-based or case-control samples when they study rare events: cancer (Breslow and Day, 1980), infant death (Currie and Neidell, 2005), consumer bankruptcy (Domowitz and Sartain, 1999), and drug trafficking (Carvalho and Soares, 2016), among many others. Case-control sampling arises frequently in biostatistics when doctors or epidemiologists study risk factors for a rare disease: random sampling may yield only a few observations with the disease among several thousands of data. In econometrics, it is often referred to as choice-based or response-based sampling because the outcome of interest is discrete choice in many economic applications (see, e.g., Chapter 6 of Manski, 2009).

Inference methods that work with random samples are generally not suitable when data are outcome-based. In the econometrics literature, parametric estimation with outcome-based samples has been investigated by Manski and Lerman (1977), Cosslett (1981), Manski and McFadden (1981), Hsieh, Manski, and McFadden (1985), Imbens (1992), and Lancaster and Imbens (1996), among others. This strand of the literature has focused mainly on the consistency or efficiency of parametric estimators in discrete response models; see e.g. McFadden (2015) for a review. In the biostatistics and epidemiology literature (e.g. Breslow, 1996), logistic regression has been the standard workhorse model in analyzing case-control studies with a more emphasis on sampling designs.

However, association between the outcome and treatment variables of interest can be measured without any parametric specification. Indeed, in case-control studies, the odds ratio has been a common measure of association, which is a practice that goes back, at least, to Cornfield (1951); in the logistic model, the logarithm of the odds ratio is equal to the regression coefficient of the treatment variable. Some authors have advocated specifying the odds ratio itself; see e.g. H. Chen (2007) and Tchetgen Tchetgen (2013).

The odds ratio is particularly popular and useful in case-control studies because it can be computed by using retrospective probabilities thanks to the Bayes rule.[1] Nevertheless, it has one drawback that its interpretation is not as intuitive as relative risk, i.e. a simple ratio of two conditional probabilities. This problem has been traditionally ignored by appealing to the so-called rare-disease assumption, under which the odds ratio can be shown to approximate relative risk; the quality of this approximation in practice has rarely been discussed. We illustrate all this background in section 2.3 in detail by using data from American Community Survey (ACS) 2018.

However, learning the ratio of two conditional probabilities is not sufficient to draw causal inferences, though it is useful to find association. In a lesser-known paper, Holland and Rubin (1988) adopt the potential outcome framework to illustrate how to identify the causal parameters in case-control studies: they emphasize the role of covariates and develop their analysis based on the assumption of strong ignorability. They use the counterfactual odds ratio and argue that it approximates the ratio of two potential-outcome probabilities of interest under the rare-disease assumption. Their work is the starting point of this paper. We share their motivation that "retrospective studies are never randomized" and develop new identification results from the perspective of partial identification (see, e.g., Manski, 1995, 2003, 2009; Tamer, 2010; Molinari, 2020, among others), without resorting to strong ignorability, nor to the rare-disease assumption.

Specifically, we adopt the assumptions of monotone treatment response (Manski, 1997, MTR hereafter) and monotone treatment selection (Manski and Pepper, 2000, MTS hereafter) and show that those assumptions enable us to bound the causal parameters in a meaningful way.[2] The MTR and MTS assumptions and related notions of monotonicity have been used in e.g. Bhattacharya, Shaikh, and

[1] *Retrospective probabilities* refer to conditional probabilities of the treatment assignment given the outcome status. Those of the outcome status given the treatment assignment will be called *prospective probabilities*.

[2] The resulting bounds turn out to be quite comparable with those under strong ignorability. See sections 3 and 6 for more details.

Vytlacil (2008, 2012), Kreider, Pepper, Gundersen, and Jolliffe (2012), Okumura and Usui (2014), Kim, Kwon, Kwon, and Lee (2018), Machado, Shaikh, and Vytlacil (2019), and Jun and Lee (2019) among others.

Our framework consists of a tripartite collection of random variables. We will write $(Y^*(1), Y^*(0), T^*, X^*)$ for the potential outcomes, the treatment of interest, and the covariates. We use (Y^*, T^*, X^*) to denote the variables that would have been observed under random sampling, i.e. $Y^* = T^*Y^*(1) + (1 - T^*)Y^*(0)$. Finally, (Y, T, X) denotes the random vector whose distribution is identified in the outcome-based sample: e.g. the distribution of (T, X) given $Y = 1$ is equal to that of (T^*, X^*) given $Y^* = 1$, but the marginal distribution of X is not the same as that of X^*. We will discuss our notation and the sampling schemes we consider in more detail in section 2.

We consider two causal parameters, i.e. *causal relative and attributable risk*, but we first focus on the former, which is defined by

$$\theta(x) := \frac{\mathbb{P}\{Y^*(1) = 1 \mid X^* = x\}}{\mathbb{P}\{Y^*(0) = 1 \mid X^* = x\}}, \tag{1}$$

where $\mathbb{P}\{Y^*(0) = 1 \mid X^* = x\} > 0$ is implicit. To identify $\theta(x)$, we face two separate challenges: one results from the usual missing data problem of potential outcomes and the other stems from the fact that the researcher does not have access to (Y^*, T^*, X^*) but only to (Y, T, X).

Our contributions are twofold. First, we articulate how the causal parameter $\theta(x)$ is related with functionals of the distribution of (Y, T, X) under two different versions of outcome-based sampling schemes: i.e. the traditional case-control sampling and case-population sampling considered in Lancaster and Imbens (1996). The latter has been used to study drug trafficking (Carvalho and Soares, 2016) and mass demonstrations (Rosenfeld, 2017) among others. It turns out that the odds ratio between Y and T conditional on $X = x$ is generally a sharp upper bound for $\theta(x)$ under the MTR and MTS assumptions. This interpretation does not require strong ignorability, nor does it the usual rare-disease assumption. Therefore, our

identification analysis shows that we can provide the conventional estimand, i.e. the odds ratio in the sample, with causal interpretation from the perspective of partial identification. Sharp bounds for causal attributable risk, which is defined by

$$\theta_{\text{AR}}(x) := \mathbb{P}\{Y^*(1) = 1 \mid X^* = x\} - \mathbb{P}\{Y^*(0) = 1 \mid X^* = x\}, \tag{2}$$

are also obtained under the same assumptions. However, they do not yield an estimand that is as standard and straightforward to espouse as the odds ratio.

Second, we propose a novel estimation algorithm for the aggregated (log) odds ratio, for which we obtain an explicit form of the efficient influence function. The estimator we build is a plug-in sieve estimator (e.g. X. Chen (2007)) and achieves the semiparametric efficiency bound (e.g. Newey, 1990, 1994). Aggregation provides a useful summary of the functional parameter $\theta(\cdot)$ without imposing any functional form assumptions. Aggregation can be done by using any weight function, but we focus on using the true marginal distribution of the covariates, i.e. the distribution of X^*. In doing so we propose a simple method of causal inference to address the fact that the distribution of X^* does not always coincide with that of X under case-control studies. We also propose an inference procedure for an aggregated version of the functional parameter $\theta_{\text{AR}}(\cdot)$. We develop an estimator of the upper bound for the causal attributable risk and we show how to conduct causal inference based on that. Our methods can be easily implemented by using standard statistical packages.[3]

Finally, we showcase our methodology by applying it to two unique datasets in the literature. The first dataset is an example of case-control sampling and comes from Delavande and Zafar (2019), who conducted a survey of college students to study the determinants of their university choice in Pakistan. The second one is an example of case-population sampling and is from Carvalho and Soares (2016), who combined a unique survey of members of drug-trafficking gangs in Rio de

[3]We have created an R package that implements the proposed estimators and the accompanying causal inference procedures. It is available on the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=ciccr.

Janeiro with the Brazilian Census. Using these datasets, we address new research questions that are not examined in the original papers: (i) the causal effect of attending private school on entering a very selective university and (ii) the causal impact of not staying in school on joining a criminal gang. Strong ignorability is unlikely to hold in these settings, but the MTR and MTS assumptions are plausible for both research questions. Our results show that attending private school does not have a substantial causal effect on the chance of entering a prestigious university: the upper bound changes by only a factor of 1.07 on average for those who entered the very selective university. In contrast, dropping out of school may increase the chance of joining a criminal gang by a factor of anywhere between 1 and 15. We supplement these results with the bounds on attributable risk and confirm the main qualitative findings. By the nature of partial identification, the large upper bound does not imply that the true causal effect is necessarily large. However, conclusions of no causal relationship can be viewed as more definitive ones, provided that the underlying MTR and MTS assumptions are credible. Therefore, our methodology offers an easy-to-implement, highly robust litmus test to understand causality with observational data.

To the best of our knowledge, we are not cognizant of directly relevant papers in the literature. In fact, the recent econometrics literature on outcome-based sampling is rather sparse; however, it is an important reality that random sampling can be expensive when the outcome of interest is rare. As demonstrated by our empirical examples, rare outcomes are often of interest not only in medicine and epidemiology but also in economics and social sciences in general. The goal of this paper is to revamp outcome-based sampling from the perspective of modern econometrics. Our paper is the first paper that nonparametrically connects the three dots: outcome-based sampling, causal inference and partial identification. We provide a further discussion on how our paper is related to the existing literature in section 9.

The remainder of the paper is organized as follows. Section 2 covers preliminaries. We describe two sampling schemes, i.e. case-control sampling and case-population sampling, after which we illustrate the odds ratio and related backgrounds by using data from American Community Survey (ACS) 2018. Section 3 presents identification results under strong ignorability as well as under the monotonicity assumptions. In section 4, we discuss how to aggregate the odds ratio. In section 5.1, we derive the semiparametric efficiency bound for the aggregated log odds ratio, and in section 5.2, we describe the estimation algorithm. In section 6, we present identification results for causal attributable risk. In section 7, we describe how to carry out causal inference on relative risk and attributable risk under the MTR and MTS assumptions. Section 8 presents two empirical examples. We conclude the paper by discussing the related literature and topics for future research in section 9. The appendices include aggregation of the odds ratio without taking the logarithm, a small Monte Carlo experiment, computational algorithms that are omitted from the main text, and all the proofs.

2. Preliminaries

2.1. **Causal parameters.** Let (Y^*, T^*, X^*) be the random variables of interest, where Y^* is a binary outcome, T^* is a binary treatment, and X^* is a vector of covariates. If we collected data via random sampling from the population of interest, we would observe (Y^*, T^*, X^*). But that is not the case under outcome-based sampling, where data consist of random draws of (T^*, X^*) given specific values of Y^*; the observed variables will be denoted by (Y, T, X) as we describe in the following subsection.

In order to discuss causal inference in this context, we introduce the usual potential outcome notation. So, $Y^*(t)$ will be the potential outcome for $T^* = t$ so that Y^* can be written as $Y^* = Y^*(1)T^* + Y^*(0)(1 - T^*)$, although the distribution of Y^* is still unidentified from the data under outcome-based sampling.

Now, we use the potential outcomes $Y^*(1)$ and $Y^*(0)$ for a causal analysis. The causal effect of T^* on the 'successful' case given $X^* = x$ can be measured by relative risk, denoted by $\theta(x)$, (or its logarithm), which is defined in equation (1).[4] Similar parameters are studied in e.g. Manski (2009, Chapter 6) in the regression context, but here the relative risk parameter is defined in terms of the potential outcomes to deal with causality.

Alternatively, one may prefer measuring the causal effect of T^* by the attributable risk $\theta_{\text{AR}}(x)$ defined in equation (2). However, it is $\theta(x)$, not $\theta_{\text{AR}}(x)$, that turns out to be closely related with the odds ratio, which has been widely used as a measure of association in case-control studies. For this reason we focus mainly on $\theta(x)$ and its relationship with the odds ratio. However, $\theta_{\text{AR}}(x)$ is still an interesting parameter in itself; it is the (conditional) average treatment effect in the binary context, which may deliver different information from $\theta(x)$. Thus, we do discuss partial identification of $\theta_{\text{AR}}(x)$ in section 6 in detail under the same assumptions and we propose easy-to-implement causal inference procedures on both attributable and relative risk in section 7.

2.2. **Bernoulli sampling.** Recall that (Y^*, T^*, X^*) are random variables that would have been observed under random sampling, but the researcher does not have access to a random sample in our setup. Instead, we assume that we have a random sample of (Y, T, X), where (Y, T, X) represent the variables that are actually observed in the sample that is drawn by the researcher's sampling design, i.e. Bernoulli sampling (e.g. Breslow, Robins, and Wellner, 2000), which we further describe and discuss below.

In Bernoulli sampling, the researcher draws a Bernoulli variable Y first from a pre-specified marginal distribution, after which she randomly draws (T, X) from $\mathcal{P}_y$ if and only if $Y = y$. So, Y is an artificial randomization device to decide from which subpopulation we will draw (T, X). If $\mathcal{P}_y$ is identical to the conditional

[4] See section 4 for using the logarithm and its relationship with the conventional log odds ratio in case-control studies.

distribution of (T^*, X^*) given on $Y^* = y$, then this is nothing but case-control sampling. Since $h_0 = \mathbb{P}(Y = 1)$ is part of the sampling scheme, we will assume that it is known.

So, the notation here distinguishes the original variables (Y^*, T^*, X^*) of interest from those (Y, T, X) in the sample; see, e.g., K. Chen (2001) and Xie, Lin, Yan, and Tang (2020) for using the same notational device. The advantage of this approach is that it becomes straightforward to apply asymptotic theory under random sampling to observations generated from (Y, T, X) because we can regard them as a collection of independent and identically distributed (i.i.d.) copies of (Y, T, X), though their joint distribution differs from that of (Y^*, T^*, X^*).

Below we will present two leading cases of Bernoulli sampling that we focus on throughout the paper. 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 $\mathcal{X}$.*

Assumption A may not be trivial in some applications. For example, if Y^* represents breast cancer and we have two covariates to consider, i.e. gender and age, then the joint support of gender and age depends highly on whether to condition on $Y^* = 1$ or not; the breast cancer population consists mostly of women. However, in this case, using the gender variable for extra stratification is appropriate. That is, we restrict ourselves to the population of women and both X^* and X represents the age; X^* is the age that would have been drawn from the population of women and X is the age that is drawn from the subpopulation of women with or without breast cancer, depending on the corresponding value of Y. Throughout the paper, we are implicit about the possibility of stratification using extra covariates (different from those included in X^*).

Let $\mathcal{P}_y(t, x) = f_{X|Y}(x \mid y)\mathbb{P}(T = t \mid X = x, Y = y)$, where $f_{X|Y}$ is the probability density (or mass) function of X given $Y = y$ for $y = 0, 1$.

Design 1 (Case-Control Sampling). *Suppose that for all* $(t,x) \in \{0,1\} \times \mathcal{X}$ *and for* $y \in \{0,1\}$,

$$f_{X|Y}(x \mid y) = f_{X^*|Y^*}(x \mid y) \text{ and } \mathbb{P}(T = t \mid X = x, Y = y) = \mathbb{P}(T^* = t \mid X^* = x, Y^* = y).$$

In other words, $\mathcal{P}_0$ *is the distribution of* (T^*, X^*) *given* $Y^* = 0$*, while* $\mathcal{P}_1$ *is that of* (T^*, X^*) *given* $Y^* = 1$.

Design 2 (Case-Population Sampling). *Suppose that for all* $(t,x) \in \{0,1\} \times \mathcal{X}$,

$$f_{X|Y}(x \mid 0) = f_{X^*}(x) \qquad \text{and } \mathbb{P}(T = t \mid X = x, Y = 0) = \mathbb{P}(T^* = t \mid X^* = x),$$
$$f_{X|Y}(x \mid 1) = f_{X^*|Y^*}(x \mid 1) \quad \text{and } \mathbb{P}(T = t \mid X = x, Y = 1) = \mathbb{P}(T^* = t \mid X^* = x, Y^* = 1).$$

In other words, $\mathcal{P}_0$ *represents the distribution of* (T^*, X^*) *of the entire population, while* $\mathcal{P}_1$ *is that of* (T^*, X^*) *conditional on* $Y^* = 1$.

Design 1 is arguably the most popular form of case-control studies and design 2, which we call *case-population sampling*, is considered in Lancaster and Imbens (1996).

In our notation the marginal distribution of (T, X) is identified from the data, while that of (T^*, X^*) may not. For instance, in design 1, we have

$$f_X(x) = f_{X^*|Y^*}(x|1)h_0 + f_{X^*|Y^*}(x|0)(1 - h_0) \neq f_{X^*}(x)$$

if $h_0 \neq \mathbb{P}(Y^* = 1)$; h_0 is part of the sampling scheme, while $\mathbb{P}(Y^* = 1)$ is the true probability of the case in the population. Further, $f_{YX}(1,x) = f_{X^*|Y^*}(x|1)h_0 = f_{X^*}(x)\mathbb{P}(Y^* = 1|X^* = x)h_0/\mathbb{P}(Y^* = 1)$, which yields the likelihood function studied in e.g. Manski and Lerman (1977). We emphasize that $\mathbb{P}(Y = 1|X = x)$ does not have economic (or structural) interpretation like $\mathbb{P}(Y^* = 1|X^* = x)$, where the latter is often modeled by rational behavior of an economic agent.

2.3. **An numerical illustration of the sampling schemes.** To further illustrate the setup of our paper, we discuss a simple example based on real data. Table 1 summarizes data from American Community Survey (ACS) 2018, cross-tabulating the

likelihood of top income by educational attainment. The sample is restricted to white males residing in California with at least a bachelor's degree.[5] The ACS sample is neither a case-control sample nor a case-population one, but we will use it here as a toy example to illustrate the key ideas of this paper. Later in section 8, we will present real-data examples of both case-control and case-population sampling.

The binary outcome 'Top Income' (Y^*) is defined to be one if a respondent's annual total pre-tax wage and salary income is top-coded.[6] The binary treatment (T^*) is defined to be one if a respondent has a master's degree, a professional degree, or a doctoral degree.

TABLE 1. Top income and education

	Beyond Bachelor's		Total
Top Income	$T^*=0$	$T^*=1$	
$Y^*=0$	10,533	6,362	16,895
$Y^*=1$	397	524	921
Total	10,930	6,886	17,816

From table 1, the proportions of top income earners are $\mathbb{P}(Y^*=1|T^*=1)\approx 0.08$ and $\mathbb{P}(Y^*=1|T^*=0)\approx 0.04$ by educational attainment. Thus, their difference and ratio are $\mathbb{P}(Y^*=1|T^*=1)-\mathbb{P}(Y^*=1|T^*=0)\approx 0.04$ and $\mathbb{P}(Y^*=1|T^*=1)/\mathbb{P}(Y^*=1|T^*=0)\approx 2.10$, respectively. In other words, going beyond a bachelor's degree is associated with a 4% increase in the likelihood of earning top incomes and doubles the chance of earning top incomes. The corresponding odds ratio is

$$\frac{\mathbb{P}(Y^*=1|T^*=1)\mathbb{P}(Y^*=0|T^*=0)}{\mathbb{P}(Y^*=0|T^*=1)\mathbb{P}(Y^*=1|T^*=0)}\approx 2.19. \tag{3}$$

[5]It is extracted from IPUMS USA (Ruggles, Flood, Goeken, Grover, Meyer, Pacas, and Sobek, 2019). The ACS is an ongoing annual survey by the US Census Bureau that provides key information about US population. The IPUMS database contains samples from the 2000-2018 ACS.

[6]In ACS 2018, the threshold income for top-coding is different across states. In our sample extract, the top-coded income bracket has median income $565,000 and the next highest income that is not top-coded is $327,000.

If we now look at the table in the *retrospective* manner, the proportions of going beyond a bachelor's degree are $\mathbb{P}(T^* = 1|Y^* = 1) \approx 0.57$ and $\mathbb{P}(T^* = 1|Y^* = 0) \approx 0.38$ by top income status. If we compute the odds ratio using these two probabilities, we have

$$\frac{\mathbb{P}(T^* = 1|Y^* = 1)\mathbb{P}(T^* = 0|Y^* = 0)}{\mathbb{P}(T^* = 0|Y^* = 1)\mathbb{P}(T^* = 1|Y^* = 0)} \approx 2.19, \tag{4}$$

which is the same as before. This is not a coincidence but a consequence of the Bayes rule, which is known as the invariance property of the odds ratio (e.g. Cornfield, 1951).

Top income is a rare event (that is, $\mathbb{P}(Y^* = 1) \approx 0.05$) in this example; if one were drawing a random sample from a population that is represented by table 1, the resulting sample would have very few observations of the top income case. Therefore, sampling from the stratification of $Y^* = 1$ is often used as a practical sampling scheme. Tables 2 and 3 are obtained by reweighting table 1 to represent a case-control sample and a case-population one, respectively.

TABLE 2. An example of case-control (design 1)

	Beyond Bachelor's		Total
Top Income	$T = 0$	$T = 1$	
$Y = 0$	527	318	845
$Y = 1$	397	524	921

Note. This table is obtained by reweighting table 1 such that $\mathbb{P}(T = t|Y = y) = \mathbb{P}(T^* = t|Y^* = y)$.

In tables 2 and 3, the conditional distributions of T given Y are related with the population of interest, but the marginal distributions of Y and T are not informative for those of Y^* and T^*. With a sample of table 2 or table 3, one cannot uncover the prospective probabilities in the population, i.e. $\mathbb{P}(Y^* = y|T^* = t)$, but the invariance property of the odds ratio can still provide us with useful information

TABLE 3. An example of case-population (design 2)

	Beyond Bachelor's		Total
Top Income	$T=0$	$T=1$	
$Y=0$	547	345	891
$Y=1$	397	524	921

Note. This table is obtained by reweighting table 1 such that $\mathbb{P}(T=t|Y=0)=\mathbb{P}(T^*=t)$ and $\mathbb{P}(T=t|Y=1)=\mathbb{P}(T^*=t|Y^*=1)$. Here, $Y=1$ represents the case of top income, but $Y=0$ is not purely the case of no top income; hence Lancaster and Imbens (1996) referred to it as a 'contaminated' control group.

through the retrospective probabilities. Specifically, if one computes the retrospective odds ratio from tables 2 and 3, we obtain the following numbers:

$$\frac{\mathbb{P}(T=1|Y=1)\mathbb{P}(T=0|Y=0)}{\mathbb{P}(T=0|Y=1)\mathbb{P}(T=1|Y=0)} \approx \begin{cases} 2.19 & \text{from table 2,} \\ 2.10 & \text{from table 3.} \end{cases} \tag{5}$$

In case of table 2, the retrospective odds ratio is 2.19, i.e. it is equal to the one we had in equation (4). As we commented earlier, this is a consequence of the Bayes rule, and it explains why the odds ratio has attracted much attention in case-control studies (Breslow, 1996). However, in case of table 3, the odds ratio is not quite the same but it is equal to the ratio of the two prospective probabilities, i.e. $\mathbb{P}(Y^*=1|T^*=1)/\mathbb{P}(Y^*=1|T^*=0) \approx 2.10$.

Causal inference based on table 1 requires assumptions, and so does the causal analysis based on table 2 or table 3. In this example, it is highly unlikely that selection into master's degree or higher is conditionally independent of top incomes. Instead of assuming strong ignorability, it might be more realistic to assume that higher degrees do not harm the chance of earning top incomes (i.e. MTR) and respondents who select higher levels of degrees are no less likely to earn top incomes, if they are randomly assigned to a different education status, than those who choose a bachelor's degree only (i.e. MTS). We start with identification in the following section.

3. Identification of Causal Relative Risk

We first introduce some notation that we will use throughout this section. For $p \in [0,1]$, define

$$r(x,p) := \begin{cases} \dfrac{p(1-h_0)\mathbb{P}(Y=1 \mid X=x)}{p(1-h_0)\mathbb{P}(Y=1 \mid X=x) + h_0(1-p)\mathbb{P}(Y=0 \mid X=x)} & \text{under design 1,} \\ \dfrac{p(1-h_0)}{h_0}\dfrac{\mathbb{P}(Y=1 \mid X=x)}{\mathbb{P}(Y=0 \mid X=x)} & \text{under design 2.} \end{cases}$$

Further let $\Pi(t \mid y,x) := \mathbb{P}(T=t \mid Y=y, X=x)$ be the retrospective binary regression function identified from the distribution of (Y,T,X). Define

$$\Gamma(x,p) := \frac{\Pi(1 \mid 1,x)}{\Pi(0 \mid 1,x)} \times \frac{\Pi(0 \mid 0,x) + r(x,p)\{\Pi(0 \mid 1,x) - \Pi(0 \mid 0,x)\}}{\Pi(1 \mid 0,x) + r(x,p)\{\Pi(1 \mid 1,x) - \Pi(1 \mid 0,x)\}},$$

where we are implicitly assuming that $\Pi(t \mid y,x) \neq 0$ for all $t,y \in \{0,1\}$.

The functions r and Γ are all identified under Bernoulli sampling because they are functionals of the joint distribution of (Y,T,X). Also, it is worth noting that

$$r(x,p_0) = \mathbb{P}(Y^*=1 \mid X^*=x) \quad \text{and} \quad r(x,0) = 0$$

under both designs 1 and 2, where $p_0 = \mathbb{P}(Y^*=1)$ is the true probability of the case in the population, which is an unidentified object. Therefore, $\Gamma(x,0)$ is nothing but the usual odds ratio, which has been popular as a measure of association in case-control studies because of its invariance property; i.e. it is equal to the prospective odds ratio in terms of (Y,T,X), and it is also equal to the odds ratio in terms of (Y^*,T^*,X^*) under design 1.

3.1. **Strong ignorability.**

The most common starting point for causal inference is probably using the strong ignorability assumptions, which consist of the following two requirements.

Assumption B (Overlap). *For all* $(t, x) \in \{0, 1\} \times \mathcal{X}$, *we have*

$$0 < \mathbb{P}\{Y^*(t) = 1 | X^* = x\} < 1 \quad \textit{and} \quad 0 < \mathbb{P}(T^* = 1 | X^* = x) < 1.$$

Assumption C (Unconfoundedness). *For all* $t \in \{0, 1\}$ *and* $x \in \mathcal{X}$,

$$\mathbb{P}\{Y^*(t) = 1 | T^* = 1, X^* = x\} = \mathbb{P}\{Y^*(t) = 1 | T^* = 0, X^* = x\}.$$

Assumption C is the key condition for strong ignorability; if we control for the covariates X^*, then the treatment assignment T^* is independent of the potential outcomes. Below we summarize the main implications and we discuss them from the perspective of partial identification.

Theorem 1. *Suppose that assumptions A to C are satisfied. Then, for all* $x \in \mathcal{X}$, $\theta(x) = \Gamma(x, p)$ *with* $p = p_0$ *under design 1 and* $p = 0$ *under design 2.*

Theorem 1 is closely related with Holland and Rubin (1988). Specifically, using design 1, Holland and Rubin (1988) shows that $\Gamma(x, 0)$ is equal to the odds ratio in terms of the potential outcomes: i.e.

$$\Gamma(x, 0) = \frac{\mathbb{P}\{Y^*(1) = 1 \mid X^* = x\}}{\mathbb{P}\{Y^*(0) = 1 \mid X^* = x\}} \frac{\mathbb{P}\{Y^*(0) = 0 \mid X^* = x\}}{\mathbb{P}\{Y^*(1) = 0 \mid X^* = x\}}$$

under strong ignorability. Theorem 1 reformulates this result in terms of relative risk, which has clearer causal interpretation, and it also extends the same idea to the case of design 2.

Theorem 1 shows point identification of $\theta(x)$ under design 2, but the case of design 1 is a bit different since p_0 is unidentified there. Traditionally, this problem has been 'ignored' by using the fact that $\Gamma(x, 0)$ is an approximation when p_0 is sufficiently close to zero by continuity; the assumption that p_0 is close to zero is often referred to as the rare-disease assumption in biostatistics and epidemiology. However, the (right-side) slope of $\Gamma(x, p)$ at $p = 0$ is generally unrestricted. Specifically,

under design 1, we have

$$\partial_p \Gamma(x,p)\big|_{p\downarrow 0} = \frac{\Pi(1 \mid 1,x)\{\Pi(1 \mid 0,x) - \Pi(1 \mid 1,x)\}}{\Pi(0 \mid 1,x)\Pi^2(1 \mid 0,x)} \frac{f_{X^*|Y^*}(x \mid 1)}{f_{X^*|Y^*}(x \mid 0)},$$

where $\Pi(t \mid y,x)$ is in fact equal to $\mathbb{P}(T^* = t \mid X^* = x, Y^* = y)$ in design 1. Therefore, p_0 being close to zero restricts neither the sign nor the magnitude of $\partial_p \Gamma(x,p)$ in a neighborhood of $p = 0$. Indeed, if T^* is a rare treatment among the non-case population, or if it is an overly common treatment among the case population, then the rare-disease approximation can be poor.[7]

However, since the function $\Gamma(x,\cdot)$ is identifiable in a case-control setup, we can provide information regarding how good or bad the rare-disease approximation will be. A general way to formalize this idea is to use the concept of partial identification. That is, if p_0 is not exactly zero but if we have some prior information on its upper bound, then the case of design 1 can be formally studied as a partial identification problem.

Assumption D. *There is a known value $\bar{p}$ such that $p_0 \in [0,\bar{p}]$.*

The idea that the researcher can place an upper bound on an unidentified object has also been used in the context of robust estimation: see e.g. Horowitz and Manski (1995, 1997). Of course, $\bar{p} = 1$ reflects the case where the researcher has no prior information about p_0 at all.

Theorem 2. *Suppose that assumptions A to D are satisfied. Under design 1, we have*

$$\min\{\Gamma(x,0),\ \Gamma(0,\bar{p})\} \leq \theta(x) \leq \max\{\Gamma(x,0),\ \Gamma(0,\bar{p})\},$$

where the bounds are sharp.

Theorem 2 follows from theorem 1 because $\Gamma(x,p)$ is monotonic in $p \in [0,\bar{p}]$; i.e. it suffices to consider the two end points to obtain sharp bounds, where one of

[7] Consider the case where there are no covariates. Suppose that smoking is fairly rare within the healthy population while it is less so among the population with lung cancer. For instance, if $\Pi(1 \mid 0) = 0.1$, $\Pi(1 \mid 1) = 0.7$, and $p_0 = 0.01$, then $\Gamma(0) - \Gamma(p_0) \approx 1.4$; i.e. when there is no causal effect, $\Gamma(p_0) = 1$, $\Gamma(0)$ is about 2.4.

the end points is the odds ratio $\Gamma(x, 0)$. In fact, if we have $Y^*(1) \geq Y^*(0)$ almost surely (i.e. the treatment is potentially beneficial but it cannot hurt) in addition to the conditions of theorem 2, then we can show that $\Gamma(x, p)$ is decreasing in p, and therefore the sharp bounds of $\theta(x)$ are given by

$$\Gamma(x, \bar{p}) \leq \theta(x) \leq \Gamma(x, 0). \tag{6}$$

In this case the odds ratio represents the maximum causal relative risk that is consistent with what is observed under design 1. Also, if there is no information for p_0, i.e. $\bar{p} = 1$, then the lower bound in equation (6) is simply one.

As the two theorems demonstrate, design 2 provides an easier environment for causal inference. It seems a bit 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.

3.2. **Beyond strong ignorability.** Strong ignorability is a popular assumption for causal inference, but it does not always deliver point identification under Bernoulli sampling. In this subsection we show that there is a set of alternative assumptions that allow endogenous treatment assignment and still imply comparable, though not the same, identifiable bounds.

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

Assumption F (Monotone Treatment Selection)**.** *For all* $t \in \{0, 1\}$ *and* $x \in \mathcal{X}$*,*

$$\mathbb{P}\{Y^*(t) = 1 \mid T^* = 1, X^* = x\} \geq \mathbb{P}\{Y^*(t) = 1 \mid T^* = 0, X^* = x\}.$$

Assumption E was first proposed by Manski (1997), while assumption F was used by Manski and Pepper (2000). Assumption E rules out the possibility of $Y^*(1) = 0$ and $Y^*(0) = 1$ in our context. For instance, if an individual, who is randomly assigned to a higher degree, does not earn high incomes, then he or she

will not be highly paid, either, when randomly assigned to no higher degree. Put differently, a higher degree is potentially beneficial, but it can never hurt.

Assumption F says that other things being equal, those who have a higher degree are at least as likely to earn high incomes, if their education attainment was randomly assigned, compared to those who did not have a higher degree. So, in substance, the treatment decision chosen by an individual reveals the 'type' of the person; those who choose to obtain a higher degree are more motivated and they would not be any less likely to earn high incomes than those who choose not to obtain a higher degree if they were randomly assigned to a different treatment status. Assumption F is trivially weaker than assumption C, and it allows individuals with 'higher ability' to self-select a higher degree.

Theorem 3. *Suppose that assumptions A, B, E and F are satisfied. Then, under both designs 1 and 2, we have*

$$1 \leq \theta(x) \leq \Gamma(x, 0),$$

where the bounds are sharp.

The sharp lower bound in theorem 3 is trivial, which follows from assumption E. The upper bound of the odds ratio is a consequence of assumption F. It is also noteworthy that the bounds in theorem 3 do not require the rare-disease assumption, or assumption D.

Theorem 3 can be compared with theorems 1 and 2. Under design 2, strong ignorability is very informative in that it ensures that the odds ratio point-identifies the causal parameter. However, under design 1, we only obtain interval identification, whether we impose strong ignorability or monotonicity; the odds ratio will be the sharp upper bound of the causal relative risk. The empirical content assumptions E and F can deliver does not depend on which of the two Bernoulli sampling schemes is used.

Therefore, with the causal parameter $\theta(x)$, or $\log\{\theta(x)\}$, in mind, the odds ratio $\Gamma(x, 0)$, or its logarithm, is a natural estimand to focus on; we discuss the practice of

taking logarithm in the following section. If the sample available is from design 2, then it is relevant how the researcher views assumption C versus assumption F. However, if the sample is from design 1, i.e. case-controlled with no 'contamination,' then it is not so much useful to impose assumption C in learning the potential maximum benefit of the treatment.

In what follows we will take theorem 3 as our basis for causal inference; using strong ignorability for causal inference is either easier or can be done similarly. From hereon we will write $\mathrm{OR}(x) = \Gamma(x, 0)$ to emphasize that this is the usual odds ratio conditional on $X = x$.

4. The odds ratio, heterogeneity, and aggregation

Since we are conditioning on a specific value of the covariate vector, theorem 3 is the most general approach to deal with potential heterogeneity in the treatment effect. However, $\log \mathrm{OR}(x)$ is generally difficult to estimate with high precision when the dimension of X^* is high.

To avoid the curse of dimensionality, it is popular in case-control studies to adopt logistic regression at the true population level: for example,

$$\mathbb{P}(Y^* = 1 | T^* = t, X^* = x) = \frac{\exp(\alpha_0 + t\alpha_1 + x^\intercal \alpha_2 + t x^\intercal \alpha_3)}{1 + \exp(\alpha_0 + t\alpha_1 + x^\intercal \alpha_2 + t x^\intercal \alpha_3)}, \tag{7}$$

which implies that

$$\log \mathrm{OR}(x) = \alpha_1 + x^\intercal \alpha_3$$

for all $x \in \mathcal{X}$ by the Bayes rule; this explains the practice of taking logarithm.[8]

The parametric assumption is convenient, but it is restrictive. For instance, the formulation in equation (7) limits the possible forms of heterogeneous causal effects; without the parametric assumption, $\log\{\mathrm{OR}(x)\}$ is generally an unknown function of x that can be highly nonlinear. Instead of parametrizing the odds ratio, we can obtain a robust summary measure by aggregating over the population

[8] Following this convention, we work with the log odds ratio from hereon, but all the results can be modified to the odds ratio without logarithm in a straightforward manner. See appendix A for details.

distribution of the covariates, i.e. $\bar{\vartheta} := \mathbb{E}\{\log \theta(X^*)\}$, where

$$0 \leq \bar{\vartheta} \leq \mathbb{E}\{\log \mathrm{OR}(X^*)\} \tag{8}$$

and the bounds are sharp under the conditions of theorem 3.[9]

However, X^* is not always observed and hence this type of aggregation is not always feasible. We start with feasible aggregation: for $y = 0, 1$, define

$$\beta(y) := \int \log \mathrm{OR}(x) dF_{X|Y}(x \mid y), \tag{9}$$

where $F_{X|Y}(\cdot \mid y)$ is the distribution of X given $Y = y$ under Bernoulli sampling. Then, $\beta(y)$ is an identified object and we have

$$\mathbb{E}\{\log \mathrm{OR}(X^*)\} = \begin{cases} (1 - p_0)\beta(0) + p_0\beta(1) & \text{under design 1,} \\ \beta(0) & \text{under design 2.} \end{cases} \tag{10}$$

Therefore, case-population sampling makes aggregation simpler as well.

5. Efficient estimation of $\beta(y)$

In this section we study semiparametrically efficient estimation of $\beta(y)$ for $y = 0, 1$. Since the mathematical structure of the likelihood function is the same for both designs 1 and 2, we do not need to distinguish them; we will simply use the generic notation using the observed variables (Y, T, X) instead of the original random variables of interest in the population, i.e. (Y^*, T^*, X^*).

5.1. The efficient influence function.

As a first step for efficient estimation, we derive the semiparametric efficiency bound. We start with the following assumptions for regularity.

Assumption G (Bounded Probabilities)**.** *There is a constant $\varepsilon > 0$ such that for each $y = 0, 1$, $\varepsilon \leq \mathbb{P}(T = 1|X, Y = y) \leq 1 - \varepsilon$ and $\varepsilon \leq \mathbb{P}(Y = 1|X) \leq 1 - \varepsilon$ almost surely.*

[9] Aggregation can be done by using a different weight function, but it seems the most natural to use the true population distribution of the covariates.

Assumption H (Regular Distribution)**.** *The distribution function $F_{X|Y}$ has a probability density $f_{X|Y}$ that satisfies $0 < f_{X|Y}(x|y) < \infty$ for all $x \in \mathcal{X}$ and $y = 0, 1$.*

Assumptions G and H are, in principle, testable since they are about the random variables observed in the sample. Therefore, assumption G is not at odds with assumption D. For example, under both designs 1 and 2, assumption G requires that $h_0 f_{X^*|Y^*}(x \mid 1)/f_{X^*}(x)$ be bounded away from zero. Assumption G is slightly stronger than what we need to derive the efficient influence function but it is useful to ensure that all the population quantities given below are well defined without spelling out all the moment conditions. Assumption H focuses on the case where X is continuous but this is only for the sake of notational simplicity; if X is discrete or mixed, then $f_{X|Y}$ should be understood as a general Radon-Nikodym density with respect to some dominating measure.

Under the Bernoulli sampling scheme, the likelihood of a single observation (Y, T, X) is given by

$$L(Y, T, X) = \big\{(1 - h_0)\mathcal{P}_0(T, X)\big\}^{1-Y}\big\{h_0\mathcal{P}_1(T, X)\big\}^{Y}, \tag{11}$$

where for $y = 0, 1$,

$$\mathcal{P}_y(T, X) = f_{X|Y}(X|y)\mathbb{P}(T = 1|X, Y = y)^T\big\{1 - \mathbb{P}(T = 1|X, Y = y)\big\}^{1-T}. \tag{12}$$

The likelihood in equation (11) is a simple mixture of two binary likelihoods. The tangent space can be derived by using regular parametric submodels $\mathcal{P}_y(T, X; \gamma)$ such that $\mathcal{P}_y(T, X; \gamma_0) = \mathcal{P}_y(T, X)$ for $y = 0, 1$. The tangent space is described in the following lemma.

Lemma 1. *Consider the Bernoulli sampling scheme of design 1 or design 2. The tangent space is given by the set of functions of the following form:*

$$s(Y, T, X) = (1 - Y)\Big[a_0(X) + \big\{T - \mathbb{P}(T = 1|X, Y = 0)\big\}b_0(X)\Big] + Y\Big[a_1(X) + \big\{T - \mathbb{P}(T = 1|X, Y = 1)\big\}b_1(X)\Big],$$

where the functions a_y and b_y are such that $\mathbb{E}\{a_y(X)|Y=y\}=0$ and $\mathbb{E}\{s^2(Y,T,X)\}<\infty$ for each $y=0,1$.

The following theorem shows that $\beta(y)$ is pathwise differentiable along the regular parametric submodels at γ_0 in the sense of Newey (1990, 1994). Before we present the theorem, define

$$w(X) := \frac{f_{X|Y}(X|0)}{f_{X|Y}(X|1)} = \frac{h_0}{1-h_0}\frac{\mathbb{P}(Y=0 \mid X)}{\mathbb{P}(Y=1 \mid X)}, \tag{13}$$

where the second equality is by the Bayes rule. Further, for $y=0,1$, define

$$\Delta_y(Y,T,X) := \frac{Y^y(1-Y)^{1-y}\{T-\mathbb{P}(T=1|X,Y=y)\}}{\mathbb{P}(T=1|X,Y=y)\{1-\mathbb{P}(T=1|X,Y=y)\}}.$$

We establish the following result using the approach taken by Hahn (1998).

Theorem 4. *Suppose that assumptions A, G and H hold and that we have a sample by Bernoulli sampling. Then, for $y=0,1$, $\beta(y)$ is pathwise differentiable and its pathwise derivative is given by*

$$F_y(Y,T,X) = \frac{Y^y(1-Y)^{1-y}}{h_0^y(1-h_0)^{1-y}}\Big\{\log \mathrm{OR}(X)-\beta(y)\Big\} - \frac{\Delta_0(Y,T,X)}{(1-h_0)w(X)^y} + \frac{w(X)^{1-y}\Delta_1(Y,T,X)}{h_0}.$$

Further, F_y is an element of the tangent space, and therefore, the semiparametric efficiency bound for $\beta(y)$ is given by $\mathbb{E}\{F_y^2(Y,T,X)\}$.

Theorem 4 shows the efficiency bound for $\beta(y)$, and it also implies that the asymptotic variance of a $\sqrt{n}$–consistent and asymptotically linear estimator of $\beta(y)$ should be $\mathbb{E}\{F_y^2(Y,T,X)\}$ by Theorem 2.1 of Newey (1994). Since $\beta(y)$ is the expectation of $\log \mathrm{OR}(X)$ with respect to the distribution of X given $Y=y$, it satisfies

$$\mathbb{E}\big\{\log \mathrm{OR}(X)-\beta(y)\big|Y=y\big\} = \mathbb{E}\left[\frac{Y^y(1-Y)^{1-y}}{h_0^y(1-h_0)^{1-y}}\Big\{\log \mathrm{OR}(X)-\beta(y)\Big\}\right] = 0, \tag{14}$$

which is the expected value of the first term that appears in $F_y(Y,T,X)$; the other terms in $F_y(Y,T,X)$ are for adjustment to address the effect of first step nonparametric estimation of $\log \mathrm{OR}(X)$ via $\mathbb{P}(T=1|X=x,Y=y)$.

5.2. **Estimation algorithms.** Efficient estimators of $\beta(y)$ for $y=0,1$ can be constructed in multiple ways. The most straightforward approach is just using equation (14), i.e. we base an estimator on

$$\beta(y) = \mathbb{E}\left\{ \frac{Y^y(1-Y)^{1-y}}{h_0^y(1-h_0)^{1-y}} \log \mathrm{OR}(X) \right\}, \tag{15}$$

where we plug in a nonparametric estimator of $\mathrm{OR}(x)$. If the resulting estimator is $\sqrt{n}$-consistent and asymptotically linear, then its asymptotic variance will be given by $\mathbb{E}\{F_y^2(Y,T,X)\}$, as we commented earlier.

Suppose that we have the sample $\{(Y_i,T_i,X_i) : i=1,\ldots,n\}$, where (Y_i,T_i,X_i)'s are i.i.d. copies of (Y,T,X). Using this sample, we propose sieve logistic estimators based on equation (15). Throughout the discussion we assume that h_0 is known since it is part of the sampling scheme. However, if it is unknown, then using $\hat{h} = \sum_{i=1}^n Y_i/n$ instead of h_0 does not change the first-order asymptotic distributions of the estimators based on (15), as long as $\mathcal{P}_0$ and $\mathcal{P}_1$ do not depend on h_0.

Recall that the odds ratio (in terms of the observed variables) is given by

$$\mathrm{OR}(x) = \frac{\mathbb{P}(T=1|X=x,Y=1)}{\mathbb{P}(T=0|X=x,Y=1)} \frac{\mathbb{P}(T=0|X=x,Y=0)}{\mathbb{P}(T=1|X=x,Y=0)},$$

where we estimate the conditional probabilities nonparametrically. Specifically, we use infinite dimensional logistic regression: i.e. for $y=0,1$,

$$\mathbb{P}(T=1|X=x,Y=y) = \frac{\exp\left\{\sum_{j=1}^\infty \phi_j(x)\mu_{j,y}\right\}}{1+\exp\left\{\sum_{j=1}^\infty \phi_j(x)\mu_{j,y}\right\}},$$

where $\{\phi_j : j=1,2,\ldots\}$ is a series of basis functions and $\{\mu_{j,y} : j=1,2,\ldots\}$ is a series of unknown coefficients for each $y=0,1$. It then follows that for each

$y = 0, 1$,

$$\log \frac{\mathbb{P}(T = 1|X = x, Y = y)}{\mathbb{P}(T = 0|X = x, Y = y)} = \sum_{j=1}^{\infty} \phi_j(x)\mu_{j,y}. \tag{16}$$

Therefore, by using equation (15) and assumption G, we obtain

$$\begin{aligned} \beta(y) &= \sum_{j=1}^{\infty} \int_{\mathcal{X}} \phi_j(x) dF_{X|Y}(x|y) \left(\mu_{j,1} - \mu_{j,0}\right) \\ &\approx \sum_{j=1}^{J_n} \int_{\mathcal{X}} \phi_j(x) dF_{X|Y}(x|y) \left(\mu_{j,1} - \mu_{j,0}\right), \end{aligned} \tag{17}$$

provided that J_n diverges to infinity as $n \to \infty$. Equation (17) suggests the following two-step sieve estimation strategy:

(i) In the first step, for each $y = 0, 1$, estimate $\{\mu_{j,y} : y = 0, 1\ j = 1, \ldots, J_n\}$ by logistic regression of T_i on $\{\phi_j(X_i) : j = 1, \ldots, J_n\}$ with the $Y_i = y$ sample.

(ii) In the second step, construct a sample analog of equation (17): i.e.

$$\widehat{\beta}(y) := \sum_{j=1}^{J_n} \int_{\mathcal{X}} \phi_j(x) d\widehat{F}_{X|Y}(x|y) \left(\widehat{\mu}_{j,1} - \widehat{\mu}_{j,0}\right), \tag{18}$$

where $\widehat{\mu}_{j,y}$'s are sieve logit estimates from the first step and

$$\int_{\mathcal{X}} \phi_j(x) d\widehat{F}_{X|Y}(x|y) = \frac{\sum_{i=1}^n Y_i^d (1 - Y_i)^d \phi_j(X_i)}{\sum_{i=1}^n Y_i^d (1 - Y_i)^d}.$$

Since the retrospective probability model is used in equation (16), we call the estimator defined in (18) the *retrospective sieve logistic estimator* of $\beta(y), y = 0, 1$. It can be computed using standard software for logistic regression, as described in algorithm 1.

The procedure described in algorithm 1 achieves the first step by running a combined logistic regression of T_i on Y_i, the sieve basis terms and the interactions between Y_i and the sieve basis terms. This is first-order equivalent since Y_i is binary and full interaction terms are included. For the second step, instead of evaluating the right-hand side of equation (18) after logistic regression, $\phi_j(X_i)$'s are demeaned

Algorithm 1: Retrospective Sieve Logistic Estimator of $\beta(1)$

Input: $\{(Y_i, T_i, X_i) : i = 1, \ldots, n\}$, tuning parameter J_n and basis functions $\{\phi_j(\cdot) : j = 1, \ldots, J_n\}$

Output: estimate of $\beta(1)$ and its standard error

1 Construct $\{\phi_1(X_i), \ldots, \phi_{J_n}(X_i) : i = 1, \ldots, n\}$, where an intercept term is excluded in ϕ_j's;

2 For each $j = 1, \ldots, J_n$, compute the empirical mean of $\phi_j(X_i)$ using only the case sample ($Y_i = 1$) and construct the demeaned version, say $\varphi_j(X_i)$, of $\phi_j(X_i)$;

3 Run a logistic regression of T_i on the following regressors: an intercept term, Y_i, $\varphi_j(X_i), j = 1, \ldots, J_n$, and interactions between Y_i and $\varphi_j(X_i), j = 1, \ldots, J_n$, using standard software;

4 Read off the estimated coefficient for Y_i and its standard error

first using only the case sample so that the resulting coefficient for Y_i is first-order equivalent to the estimator defined in equation (18). The advantage of the formulation in algorithm 1 is that the standard error of $\widehat{\beta}(1)$ can be read off directly from standard software without any further programming. It is straightforward to modify algorithm 1 for estimating $\beta(0)$. One has to compute the empirical mean of $\phi_j(X_i)$ using only the control sample ($Y_i = 0$) for the demeaning step.

Sieve logistic estimators have been popular in the literature, including the propensity score estimator used in Hirano, Imbens, and Ridder (2003). To the best of our knowledge, it is novel to adopt retrospective sieve logistic estimators in the context of case-control studies. It is not difficult to work out formal asymptotic properties of our proposed sieve estimator in view of the well-established literature on two-step sieve estimation (see, e.g., Ai and Chen, 2003, 2012; Ackerberg, Chen, Hahn, and Liao, 2014, among many others). Since this is now well understood in the literature, we omit details for brevity of the paper. Appendix B reports the results of a small Monte Carlo experiment that illustrates the finite-sample performance of the proposed estimators of $\beta(1)$ and $\beta(0)$.

6. Identification of Attributable Risk

In this section we present identification results for the alternative causal parameter $\theta_{\mathrm{AR}}(x)$ defined in equation (2) under the conditions we discussed in section 3. Define

$$\Gamma_{\mathrm{AR}}(x,p) := \frac{\Pi(1 \mid 1,x)}{\Pi(1 \mid 0,x) + r(x,p)\{\Pi(1 \mid 1,x) - \Pi(1 \mid 0,x)\}} - \frac{\Pi(0 \mid 1,x)}{\Pi(0 \mid 0,x) + r(x,p)\{\Pi(0 \mid 1,x) - \Pi(0 \mid 0,x)\}},$$

where $\Pi(t \mid y,x)$ and $r(x,p)$ are defined in section 3. Note that $\Gamma_{\mathrm{AR}}(x,0)$ is not the odds ratio anymore but is the difference of two probability ratios.

Theorem 5. *Suppose that assumptions A to C are satisfied. Then, for all $x \in \mathcal{X}$, $\theta_{\mathrm{AR}}(x) = r(x,p_0)\Gamma_{\mathrm{AR}}(x,p)$ with $p = p_0$ under design 1 and $p = 0$ under design 2.*

Theorem 5 considers the case of strong ignorability. Unlike $\theta(x)$, $\theta_{\mathrm{AR}}(x)$ depends on p_0 in both design 1 and design 2. Further, the rare-disease approximation does not provide anything useful because if $p_0 \approx 0$, then $r(x,p_0) \approx r(x,0) = 0$ by continuity. However, the partial identification approach is still valid.

Theorem 6. *Suppose that assumptions A to D are satisfied. Then, for all $x \in \mathcal{X}$, the sharp identifiable bounds of $\theta_{\mathrm{AR}}(x)$ are given by the intervals*

$$\begin{cases} \left[\min\limits_{p\in[0,\bar{p}]} r(x,p)\Gamma_{\mathrm{AR}}(x,p),\ \max\limits_{p\in[0,\bar{p}]} r(x,p)\Gamma_{\mathrm{AR}}(x,p)\right] & \text{under design 1,} \\ \left[\min\limits_{p\in[0,\bar{p}]} r(x,p)\Gamma_{\mathrm{AR}}(x,0),\ \max\limits_{p\in[0,\bar{p}]} r(x,p)\Gamma_{\mathrm{AR}}(x,0)\right] & \text{under design 2.} \end{cases}$$

Theorem 6 is a simple corollary of theorem 5. If the aggregated parameter $\bar{\theta}_{\mathrm{AR}} := \int_{\mathcal{X}} \theta_{\mathrm{AR}}(x)\omega(x)ds$ for some weight function ω, then the sharp bounds can be obtained by taking max/min over p_0 after aggregation.

Unlike the case of $\theta(x)$, strong ignorability does not lead to point identification under design 2. Further, similarly to the case of $\theta(x)$, the bounds under strong ignorability are quite comparable to those under the monotonicity assumptions as

we show in the following theorem. Therefore, the empirical content of the strong ignorability assumptions is limited as before.

Theorem 7. *Suppose that assumptions A, B and D to F are satisfied. Then, for all $x \in \mathcal{X}$, we have*

$$0 \leq \theta_{\mathrm{AR}}(x) \leq \begin{cases} \max_{p \in [0,\bar{p}]} r(x,p)\Gamma_{\mathrm{AR}}(x,p) & \text{under design 1,} \\ \max_{p \in [0,\bar{p}]} r(x,p)\Gamma_{\mathrm{AR}}(x,0) & \text{under design 2,} \end{cases}$$

where the bounds are sharp.

The sharp upper bounds in theorem 7 are the same as the one presented in theorem 5. Therefore, replacing strong ignorability with assumptions E and F does not cost anything in terms of the sharp upper bound of $\theta_{\mathrm{AR}}(x)$. This result is consistent with the case of $\theta(x)$. Also, similarly to the case of $\theta(x)$, the trivial bound of 0 in theorem 7 is a consequence of assumption E.

7. Causal Inference on Relative Risk and Attributable Risk

In this section, we describe how to carry out causal inference on relative risk and attributable risk under the MTR and MTS assumptions. First, regarding relative risk, we propose to use equation (10) as a basis of causal inference. Specifically, under the MTR and MTS, we consider the upper bound on $\exp[\mathbb{E}\{\log \theta(X^*)\}]$:[10]

$$\exp[\mathbb{E}\{\log \theta(X^*)\}] \leq \begin{cases} \exp\{p_0\beta(1) + (1-p_0)\beta(0)\} & \text{under design 1,} \\ \exp\{\beta(0)\} & \text{under design 2.} \end{cases} \tag{19}$$

Causal inference under design 2 follows immediately from section 5 since the $(1-\alpha)$ one-sided confidence interval for $\exp\{\beta(0)\}$ can be constructed by $\left[1, \exp\{\widehat{\beta}(0) + z(1-\alpha)\widehat{s}(0)\}\right]$, where $\widehat{s}(0)$ is the standard error of $\widehat{\beta}(0)$, $z(1-\alpha) := \Phi^{-1}(1-\alpha)$ and $\Phi(\cdot)$ is the standard normal distribution function. Therefore, we focus on the case of design 1 below.

[10] Note that $\exp[\mathbb{E}\{\log \theta(X^*)\}] \leq \mathbb{E}\{\theta(X^*)\}$ by Jensen's inequality. We take $\exp[\mathbb{E}\{\log \theta(X^*)\}]$ as a summary measure for relative risk because an average of the logarithm of relative risk is less likely to be affected unduly by outliers than that of relative risk itself. See Appendix A for further discussions.

Under design 1 the right-hand side on equation (19) is identifiable up to the unknown p_0. Therefore, we can conduct causal inference conditional on p_0. In view of this we propose constructing a confidence band for $\tilde{\beta}(p) := p\beta(1) + (1-p)\beta(0)$ that is uniform in $p \in [0,1]$.[11] This can be done in the following way.

Let $u(1-\alpha) := z(1-\alpha/2)\max\{\widehat{s}(1), \widehat{s}(0)\}$, where $\widehat{s}(j)$'s are the standard errors that we obtain from the algorithm described in section 5.2. That is, $u(1-\alpha)$ is the usual two-sided normal critical value times the maximum of the standard errors of $\widehat{\beta}(1)$ and $\widehat{\beta}(0)$. We then have

$$\mathbb{P}\big[\forall p \in [0,1],\ \exp\{\tilde{\beta}(p)\} \le \exp\{p\widehat{\beta}(1) + (1-p)\widehat{\beta}(0) + u(1-\alpha)\}\big] \ge 1-\alpha. \quad (20)$$

The asymptotic coverage rate in equation (20) is conservative. Achieving the asymptotically exact rate requires that we use the limiting distribution of $\min\{\widehat{\beta}(1) - \beta(1),\ \widehat{\beta}(0) - \beta(0)\}$, which is not normal and is not readily available from the estimation algorithm given in section 5.2. Equation (20) can be verified as follows. Let $\bar{\beta}(j) = \widehat{\beta}(j) - \beta(j)$ for $j = 0, 1$. Then, equation (20) follows from

$$\begin{aligned}
&\mathbb{P}\big[\inf_{p\in[0,1]}\{p\bar{\beta}(1) + (1-p)\bar{\beta}(0)\} \le -u(1-\alpha)\big] \\
&= \mathbb{P}\{\bar{\beta}(1) \le -u(1-\alpha),\ \bar{\beta}(1) < \bar{\beta}(0)\} + \mathbb{P}\{\bar{\beta}(0) \le -u(1-\alpha),\ \bar{\beta}(1) \ge \beta(0)\} \\
&\le \mathbb{P}\{\bar{\beta}(1) \le -u(1-\alpha)\} + \mathbb{P}\{\bar{\beta}(0) \le -u(1-\alpha)\} \le \alpha/2 + \alpha/2.
\end{aligned}$$

We now turn to causal inference on attributable risk. The identification results in section 6 implies that the MTR and MTS assumptions lead to

$$\mathbb{E}\{\theta_{\mathrm{AR}}(X^*)\} \le \begin{cases} (1-p_0)\beta_{\mathrm{AR}}(p_0, 0) + p_0\beta_{\mathrm{AR}}(p_0, 1) & \text{under design 1,} \\ \mathbb{E}\{r(X, p_0)\Gamma_{\mathrm{AR}}(X, 0) \mid Y = 0\} & \text{under design 2,} \end{cases} \quad (21)$$

where $\beta_{\mathrm{AR}} : [0,1] \times \{0,1\} \to \mathbb{R}$ is defined by

$$\beta_{\mathrm{AR}}(p, y) := \mathbb{E}\{r(X, p)\Gamma_{\mathrm{AR}}(X, p) \mid Y = y\}. \quad (22)$$

[11] The procedure can be modified if assumption D with $0 < \bar{p} < 1$ is adopted.

Now, the case of design 2 is not immediate, though it is still simpler than the case of design 1. The right-hand sides on equation (21) are identifiable up to the unknown value p_0 again. Therefore, for a given $p_0 = p$, it is straightforward to estimate the right-hand sides on equation (21) by sample-analog estimators.

To appreciate the form of the upper bounds in equation (21), we first look at the simpler case of design 2: i.e. $\mathbb{E}\{r(X,p)\Gamma_{\text{AR}}(X,0) \mid Y=0\} = p\xi_{\text{CP}}$, where

$$\xi_{\text{CP}} := \frac{1-h_0}{h_0}\mathbb{E}\left[\frac{\mathbb{P}(Y=1 \mid X)}{\mathbb{P}(Y=0 \mid X)}\left\{\frac{\Pi(1 \mid 1,X)}{\Pi(1 \mid 0,X)} - \frac{\Pi(0 \mid 1,X)}{\Pi(0 \mid 0,X)}\right\}\middle| Y=0\right]. \tag{23}$$

Unlike the upper bound on relative risk in equation (19), the upper bound on attributable risk depends on $\mathbb{P}(Y=1 \mid X)$; therefore, it is necessary to specify a model for $\mathbb{P}(Y=1 \mid X=x)$ in addition to the retrospective probability $\Pi(t \mid y,x)$.

Suppose that we have constructed fitted values $\{\widehat{\mathbb{P}}(Y=1 \mid X_i) : i=1\ldots,n\}$ and $\{\widehat{\Pi}(t \mid y,X_i) : t=0,1,\ y=0,1,\ i=1\ldots,n\}$ by using parametric or sieve logit estimators. Then, we can estimate ξ_{CP} by

$$\widehat{\xi}_{\text{CP}} := \left(\sum_{i=1}^n Y_i\right)^{-1}\sum_{i=1}^n (1-Y_i)\left[\frac{\widehat{\mathbb{P}}(Y=1 \mid X_i)}{\widehat{\mathbb{P}}(Y=0 \mid X_i)}\left\{\frac{\widehat{\Pi}(1 \mid 1,X_i)}{\widehat{\Pi}(1 \mid 0,X_i)} - \frac{\widehat{\Pi}(0 \mid 1,X_i)}{\widehat{\Pi}(0 \mid 0,X_i)}\right\}\right],$$

where h_0 is replaced with the sample mean of Y_i. In principle, it is possible to derive the asymptotic distribution of $\sqrt{n}(\widehat{\xi}_{\text{CP}} - \xi_{\text{CP}})$ and develop a sample analog estimator of the resulting asymptotic variance; however, it would be tedious to work out details. Alternatively, as is standard in empirical practice, we may use the bootstrap instead. Specifically, the $(1-\alpha)$ uniform confidence band has the form[12]

$$p \in [0,1] \mapsto \left[0,\ p \cdot u_{\text{AR,CP}}(1-\alpha)\right],$$

where $u_{\text{AR,CP}}(1-\alpha)$ is the one-sided critical value obtained by nonparametrically bootstrapping the distribution of $\sqrt{n}(\widehat{\xi}_{\text{CP}} - \xi_{\text{CP}})$; since p shows up in a simple linear form, there is no need to modify the critical value $u_{\text{AR,CP}}(1-\alpha)$ for uniformity. Asymptotic validity of the nonparametric bootstrap in two-step semiparametric

[12]The researcher may adopt assumption D with $0 < \bar{p} < 1$ by focusing on domain $[0,\ \bar{p}]$.

models has been well studied (e.g. Chen, Linton, and Van Keilegom, 2003). In practice, the naïve bootstrap may suffer from a finite sample bias problem because the ratios of probabilities are estimated and they are averaged across observations. To mitigate this issue, we recommend using Efron (1982)'s bias-corrected one-sided percentile interval for ξ_{CP} to obtain $u_{\mathrm{AR,CP}}(1-\alpha)$.

We now consider design 1. Although the form of $r(X,p)\Gamma_{\mathrm{AR}}(X,p)$ is more involving under case-control sampling, a plug-in estimator can be constructed again by suitable sample averages based on equation (22), using estimates of both $\mathbb{P}(Y=1 \mid X=x)$ and $\Pi(t \mid y,x)$. However, in order to obtain a uniform confidence band in this case, it would be necessary to derive an asymptotic linear expansion and obtain a feasible version of the corresponding leading term; therefore, for simplicity, we simply propose to use one-sided pointwise confidence intervals using Efron's bias-corrected percentile intervals. Computational details regarding causal inference on attributable risk are given in Appendix C.

8. Examples: case-control and case-population sampling

8.1. **Attending private school and entering a very selective university.** Delavande and Zafar (2019) studied the determinants of students' university choice in Pakistan by conducting a survey with male students enrolled in different types of universities: two Western-style universities, one Islamic university, and four madrassas, all located in Islamabad/Rawalpindi and Lahore. As they first selected universities before conducting the survey and then sampled students within each university, their sample is not randomly drawn from the population of university students.

We focus on the subsample of their data, namely those enrolled in the two Western-style universities. For these 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). In our empirical

exercise, we restrict the population to those who entered either VSU or SU and define the binary outcome to be whether a student entered VSU. Because the outcome variable refers to which university a student entered, the subsample we consider can now be regarded as a case-control sample, i.e. our design 1. That is, the case students correspond to those enrolled in VSU and the control students mean those who enrolled in SU, respectively.

The binary treatment we consider is whether a student *attended private school before university*. In other words, we are interested in quantifying the causal effect of attending private school on the chance of entering VSU.

TABLE 4. University entrance and private school

	Private School		Total
University	$T = 0$	$T = 1$	
$Y = 0$ (SU)	151	332	483
$Y = 1$ (VSU)	51	155	206
Total	202	487	689

Table 4 shows the likelihood of entering VSU by private school attendance before university. The empirical odds ratio is 1.38. The unconfoundedness assumption is unlikely to hold in this example since those who attended private school before university are likely to have more resourceful parents who could help their children enter a more selective university. This concern may not go away even if we condition on parental income and wealth since it is likely that unobserved parental abilities and resources could affect their children's university choice. However, the MTR and MTS assumptions are more plausible since (i) private school is likely to be no inferior input to university preparations (hence, MTR) and (ii) it is probable that those who actually attended private school have more resources for university preparations than those who did not attend private school (thus, MTS). Under the MTR and MTS, the odds ratio of 1.38 is the sharp upper bound on causal relative risk. That is, attending private school may not increase the causal probability of entering VSU at all or may increase it by a factor of anywhere between 1 and

1.38. This suggests that the effect of attending private school is, at best, modest excluding large impacts.

We now try to control for family background, which is likely to be an important determinant of university entrance. Table 5 reports sample averages of private school attendance and a number of family background variables by the outcome variable.[13] On average, parents of VSU students were more educated, earned higher incomes and had more wealth, measured by the ownership of home, car and so on.

TABLE 5. Summary statistics of the case-control sample

	VSU	SU
Attended private school before university	0.75	0.69
At least one college-educated parent	0.89	0.67
Parents' monthly income	193.38	95.89
Above median income	0.93	0.80
Parents own home	0.92	0.87
Parents own television	0.91	0.84
Parents own cell phone	0.92	0.80
Parents own computer	0.84	0.70
Parents own car	0.84	0.67
Sample size	206	483

Notes. Each entry shows the sample mean. Parents' monthly income is in thousands of rupees and all other variables are binary indicator variables. The case and control samples consist of those who entered a very selective university (VSU) and a selective university (SU), respectively.

We control for parental background by including the following covariates: (i) six binary indicator variables whether at least one parent is college-educated, parents own home, television, cell phone, computer, and car, and (ii) cubic b-splines terms of parents' monthly income with three inner knots. This corresponds to a partially linear specification with a sieve approximation on parents' income.

[13]Table 5 basically replicates Table 1 of Delavande and Zafar (2019). The minor difference is that two observations in SU are excluded in table 5 because of missing information on parental income.

TABLE 6. Estimation results of attending private school on entering VSU

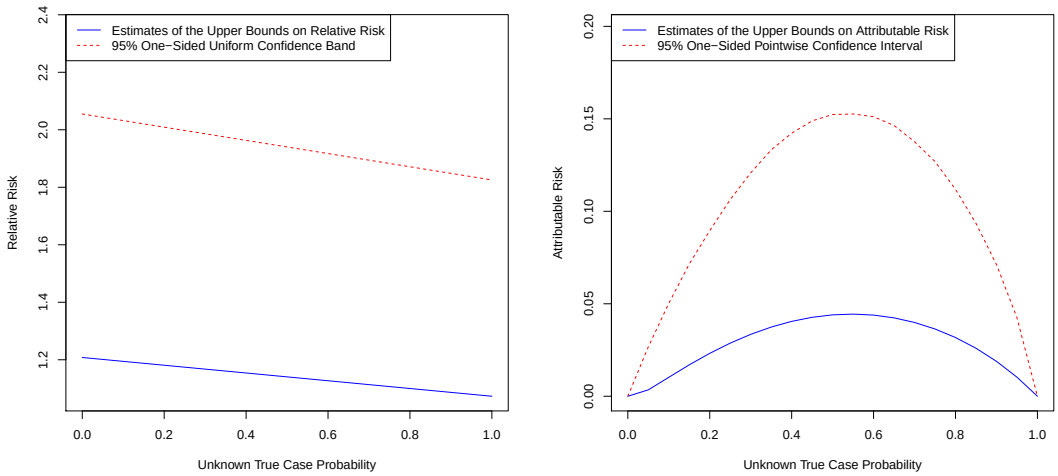
	(1) Case $(y = 1)$ VSU	(2) Control $(y = 0)$ SU
$\beta(y)$	0.07	0.19
95% confidence interval	$[0, 0.48]$	$[0, 0.63]$
$\exp[\beta(y)]$	1.07	1.21
95% confidence interval	$[1, 1.61]$	$[1, 1.89]$

Note. Parental background is controlled for when fitting retrospective binary logistic regression models. See the main text for details.

Table 6 reports estimation results. The retrospective sieve estimate of $\beta(1)$ is 0.07, which is smaller than that of $\beta(0)$, thereby suggesting that there might be heterogeneity among individuals. However, both estimates are estimated imprecisely probably due to the relatively small sample size, which indicates that the difference between the two estimates might be driven by sampling uncertainty. We also present point estimates of $\exp\{\beta(y)\}, y = 0, 1$ and their confidence intervals under the scenario that the point estimate is the upper bound on causal relative risk because the MTR and MTS assumptions are more plausible than strong ignorability in this example. The estimates of $\exp\{\beta(y)\}$ are comparable to the usual odds ratio in terms of its scale; therefore, they can be interpreted similarly. For example, 1.07 of $\exp\{\widehat{\beta}(1)\}$ roughly means that attending private school increases the upper bound for the chance of entering VSU by a factor of only 1.07. The end point of the 95% confidence interval ranges from 1.61 to 1.89, which includes the unconditional odds ratio of 1.38.

We now illustrate our methods in section 7 for conducting causal inference on relative risk and attributable risk under the MTR and MTS assumptions. We obtain the upper bounds on causal relative risk using the estimates of $\beta(1)$ and $\beta(0)$ described above. For attributable risk, we estimate the prospective model $\mathbb{P}(Y = 1 \mid X)$ using a linear logit model with only two covariates: an indicator for at least one college-educated parent and parents' monthly income; analogously,

FIGURE 1. Causal inference on relative risk and attributable risk: bounding the effects of attending private school on entering VSU



Note. The left panel shows the estimates and the 95% one-sided uniform confidence band for the upper bounds on relative risk and the right panel the estimates and the 95% one-sided pointwise confidence intervals for the upper bounds on attributable risk, as functions of the unknown true case probability, i.e., $\mathbb{P}(Y^* = 1)$.

we specify the retrospective model $\Pi(t \mid y, x)$ using a linear logit model with Y interacting with the same two covariates. We chose a less demanding specification here because the estimand for attributable risk is much more complex and causal inference is now based on the bootstrap. The number of bootstrap replications was 10,000.

Using the dataset from Delavande and Zafar (2019) alone, the true value of $p_0 = \mathbb{P}(Y^* = 1)$ is unidentifible; however, as we can see from figure 1, we can trace out the upper bounds as functions of unknown true case probability, thereby providing a tool for causal inference. We first analyze the left panel of figure 1. If we take the point estimate at face value, the impact of attending private school increases the probability of entering VSU by a factor of at most 1.21. Furthermore, based on the uniform confidence band, it is highly unlikely that attending private school increases the probability of being enrolled in VSU by a factor of more than 2. We now look at the right panel of figure 1. By definition, attributable

risk must be 0 when $p = 0$ or $p = 1$ since there will be no difference between $\mathbb{P}\{Y^*(1) = 1|X = x\}$ and $\mathbb{P}\{Y^*(0) = 1|X = x\}$ in either case. Thus, under the MTR and MTS assumptions, the upper bound is approximately inverted U-shaped. The maximum point estimate of the upper bounds is 0.044 at $p = 0.55$ and the maximum value of the confidence intervals is 0.153. In words, it is very unlikely that private school increases the chance of entering VSU by 15 percent.

Using this example, we have demonstrated that it is unnecessary to assume strong ignorability nor the rare-disease assumption to infer the causal effect of pre-university school decisions on the university choice. Our estimates are indicative of a relatively small positive effect of attending private school on entering a very selective university, if it exists at all. Our results are in line with the literature on school vouchers that finds that access to private schools often have modest effects on children's outcomes (see, e.g., Epple, Romano, and Urquiola, 2017; MacLeod and Urquiola, 2019).

8.2. **Not attending school and joining a gang.** In this section, we revisit Carvalho and Soares (2016) and use their dataset to illustrate the usefulness of our approach. They combine a unique survey of members of drug-trafficking gangs in favelas (slums) of Rio de Janeiro with the 2000 Brazilian Census. Hence, it is an empirical example of case-population sampling considered in Lancaster and Imbens (1996), i.e. our design 2. Among other things, they estimate a model of selection into the gang, which is basically a prospective binary regression model, using the generalized method of moments (GMM) estimator proposed by Lancaster and Imbens (1996). 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" (see Table 4 in Carvalho and Soares (2016) for estimation details).

To estimate their model, Carvalho and Soares (2016) use 5 covariates: race, age, illiteracy, house ownership (a measure of family wealth), and religiosity. They

TABLE 7. Summary Statistics of the Case-Population Sample

	(1) Case Gang members	(2) Population Men aged 10-25
Not in school	0.901	0.458
Black	0.269	0.142
Age	16.722	17.526
Illiterate	0.094	0.041
Owns house	0.735	0.832
No religion	0.426	0.237
Sample size	223	17175

Notes. Each entry shows the sample mean. Age is in years and all other variables are binary indicator variables. The population consists of men aged 10-25 living in Rio's Favelas.

comment that they "focus on characteristics that are more likely to be predetermined" and that "years of schooling may be endogenous to entry - i.e., joining the gang may lead members to drop out of school." Indeed, 90 percent of gang members are not in school, whereas 46 percent of men aged 10-25 are not in school. Table 7 provides summary statistics of the estimation sample.

TABLE 8. Estimation results of currently not attending school on relative risk of joining a gang

	(1) Case Gang members	(2) Population Men aged 10-25
$\beta(y)$	2.90	2.71
95% confidence interval	[0, 3.36]	[0, 3.19]
$\exp[\beta(y)]$	18.10	15.01
95% confidence interval	[1, 28.90]	[1, 24.39]

Note. Race, age, illiteracy, house ownership, and religiosity are linearly controlled for when fitting retrospective binary logistic regression models.

In this section, we regard *currently not attending school* as the treatment variable. Clearly, the unconfoundedness assumption is implausible here; however, the MTR and MTS assumptions are more plausible: not being in school may increase the chance of joining a gang (hence, MTR) and those who are not in school can have a higher counterfactual probability of becoming a gang member than those who are actually in school (thus, MTS).

Table 8 presents estimation results under the MTR and MTS assumptions. In retrospective binary logistic regression models, we control for the same five covariates linearly as in Carvalho and Soares (2016). Estimates of $\beta(y)$ and $\exp\{\beta(y)\}$ along with their one-sided confidence intervals are reported for $y = 1$ in column (1) and $y = 0$ in column (2), respectively. In case-population sampling (design 2) under MTR and MTS, the sharp upper bound on $\mathbb{E}\{\log\theta(X^*)\}$ is

$$0 \leq \mathbb{E}\{\log\theta(X^*)\} \leq \mathbb{E}\{\log \mathrm{OR}(X^*)\} = \beta(0),$$

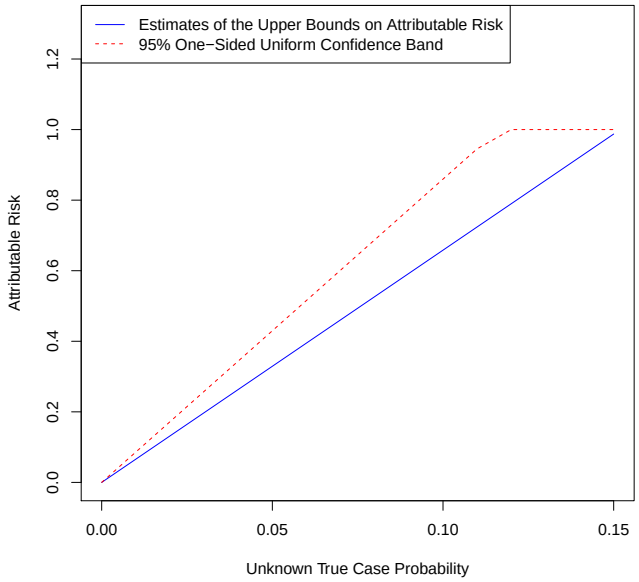
whose point estimate is 2.71 with the 95% one-sided confidence interval of $[0, 3.19]$. To be more in line with relative risk, we consider

$$\exp[\mathbb{E}\{\log\theta(X^*)\}] \leq \exp\{\beta(0)\},$$

whose point estimate is 15.01 with the 95% one-sided confidence interval of $[0, 24.39]$. Roughly speaking, our estimates suggest that those who are not in school can have relative risk of joining a gang anywhere between 1 and 15 times (in terms of the point estimate) or 24 times (in terms of the confidence interval). Our estimates do not mean that the causal impact of not being in school is necessarily large, but they imply that we cannot exclude confidently the possibility that not currently attending school could cause the relative risk of joining a gang to increase by a factor of 24.

We now supplement the results on relative risk with causal inference on attributable risk under the MTR and MTS assumptions. For attributable risk, we estimate

FIGURE 2. Causal inference on attributable risk: bounding the effects of currently not attending school on joining a gang



Note. The figure shows the estimates and the 95% one-sided uniform confidence band for the upper bounds on attributable risk, as functions of the unknown true case probability, i.e., $\mathbb{P}(Y^* = 1)$.

the prospective model $\mathbb{P}(Y = 1 \mid X)$ using a linear logit model with the same covariates as above; analogously, we specify the retrospective model $\Pi(t \mid y, x)$ using a linear logit model with Y interacting with the same covariates. The domain of p was set $[0, \bar{p}]$ with $\bar{p} = 0.15$, which includes three values of p (that is, 0.05, 0.10, 0.15) considered in Carvalho and Soares (2016). The number of bootstrap replications was 1,000. The estimated upper bounds are large: 0.33 at $p = 0.05$, 0.66 at $p = 0.10$ and 0.99 $p = 0.15$. The uniform confidence band is truncated at 1 since attributable risk cannot be larger than 1.

Overall, our results are, at best, 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. In short, the empirical exercise in this section points out two potential advantages of using our approach. First, it is possible

to consider endogenous treatment, provided that the MTR and MTS assumptions are plausible; second, it is unnecessary to assume a particular value of the true case probability, which is a required input to the GMM estimator of Lancaster and Imbens (1996) and other econometric methods for case-control studies.

9. Related Literature and Future Research

The literature on causal inference using observational data is vast and the literature on non-random sampling is extensive. We conclude in this section by discussing some of the important papers in the context of what we have achieved in this paper.

We have labeled designs 1 and 2 together as Bernoulli sampling, which is the term that we borrowed from Breslow, Robins, and Wellner (2000). The two sampling schemes have been studied under different names by other authors. For instance, Imbens and Lancaster (1996) refer to design 1 as multinomial sampling, and Lancaster and Imbens (1996) call design 2 case-control sampling with contamination, which is borrowed from Heckman and Robb (1985).

The objective of Heckman and Robb (1985) is to estimate the impact of training on earnings under various data scenarios. In that study they discuss common data problems such as oversampling of trainees or "contamination" in the control group, i.e. the training status of the individuals in the control group being unknown. Although the sampling schemes of Heckman and Robb (1985) are similar to designs 1 and 2, they are distinct in the sense that they are not outcome-based but *treatment-based* sampling. In our context, having a control group drawn from the whole population without conditioning on the outcome status makes it easier, not harder, to identify the causal relative risk parameter. For this reason we have referred to design 2 as *case-population* sampling in order to remove connotations of negativeness from the word "contamination."

Estimating the average treatment effect under treatment-based sampling has been studied by other authors as well. For instance, Heckman and Todd (2009)

point out that a matching estimator can be implemented by using the odds ratio of the propensity score fit on the sample because it is a monotone transformation of the true propensity scores. Kennedy, Sjölander, and Small (2015) show that one can estimate the average treatment effect on the treated without the knowledge of the true population probability of the treatment. Assuming the latter is known, Hu and Qin (2018) and Zhang, Hu, and Liu (2019) have developed weighted estimators of the average treatment effect. However, all these methods are based on strong ignorability, and to the best of our knowledge, we are not aware of any work that does not rely on it. We leave it for future research how to extend the approach taken in this paper to the context of treatment-based sampling.

The term Bernoulli sampling has been alternatively used by e.g. Kalbfleisch and Lawless (1988) to describe the case where an individual unit is randomly drawn from the entire population but it is retained or discarded with stratum-specific probabilities. Imbens and Lancaster (1996) use the same terminology, while they call our design 1 multinomial sampling as we mentioned earlier. The case where a given number of observations are randomly drawn from each stratum has been traditionally called the classical stratified sampling scheme (e.g. Hausman and Wise, 1981). However, Imbens and Lancaster (1996) have shown that there is no meaningful difference among the three schemes in that they lead to the same likelihood function to estimate the parameters that appear in the choice probabilities. Since this paper is concerned about a binary outcome, Bernoulli sampling seems more appropriate than multinomial sampling.

In the literature on choice-based sampling, the objective is usually efficiently estimating the parameters that appear in the parametrically specified prospective probabilities. Manski and Lerman (1977) propose a weighted likelihood approach for this purpose under outcome-based sampling. Cosslett (1981) shows that it is feasible to compute the full maximum likelihood estimator. By far the most common specification is the logistic model. However, as Xie and Manski (1989) point out, the logit model can be quite misleading under outcome-based sampling, if

the truth is not logistic. Despite its convenience, the logistic specification imposes restrictions on the form of heterogeneity in the causal effect. In contrast, our approach does not restrict the shape of the causal relative risk function $\theta(\cdot)$, thereby allowing an unrestricted form of heterogeneity in the causal treatment effect.

Many papers in this literature use the term "semiparametric" to describe the fact that the marginal distribution of the regressors are left unspecified in their analysis, while the prospective probability, i.e. the conditional distribution of the outcome given the regressors, is still parametric: see e.g. Imbens and Lancaster (1996) and Breslow, Robins, and Wellner (2000). By contrast, our approach is semi-nonparametric in the sense of X. Chen (2007) because we do not impose parametric restrictions anywhere. Instead of relying on the parametric assumption, we directly target the aggregated log odds ratio as the estimand of interest, we articulate its relationship with the fundamental causal parameter of interest, and we have derived the efficiency bound for the estimand under Bernoulli sampling. By combining all these results we can draw robust and efficient inferences on the causal parameter of interest.

In the statistics and epidemiology literature, misspecification and robustness has been addressed from a different perspective. For instance, H. Chen (2007) considers estimating the parameters that appear in the odds ratio in such a way that consistency and asymptotic normality follows as long as either the prospective or the retrospective probability is correctly specified: this approach is known as a doubly robust estimation method. Tchetgen Tchetgen, Robins, and Rotnitzky (2010) take a similar approach, but their estimator is simpler to implement than H. Chen (2007)'s; it is then further operationalized by Tchetgen Tchetgen (2013) under the finite-dimensional logistic assumption. It is also noteworthy that statisticians and epidemiologists have maintained an active research agenda in case-control studies unlike econometricians. In addition to the aforementioned papers, for instance, Zhou, Herring, Bhattacharya, Olshan, Dunson, and Study (2016) investigate how

to deal with high dimensional predictors in the case-control setup using a non-parametric Bayesian approach.

Causal inferences in case-control studies are not a new topic. As we mentioned earlier, Holland and Rubin (1988) use a similar framework to ours, where they relate the retrospective odds ratio with the odds ratio of the potential outcomes. More recently, Månsson, Joffe, Sun, and Hennessy (2007) discuss the issues and problems with using the propensity scores in case-control studies. Rose (2011) presents a comprehensive review on this topic and she proposes a method of causal inference based on the assumption that the true p_0 is known by a prior study. See also Van der Laan and Rose (2011).

Finally, we have focused on the causal parameter defined by a ratio, but it is probably fair to say that a difference (i.e. causal attributable risk in the binary setup) is a more common measure in econometrics (e.g. Hahn, 1998; Hirano, Imbens, and Ridder, 2003). We present identification results for both relative risk and attributable risk in the paper; it turns out that the ratio is mathematically more convenient under outcome-based sampling thanks to the invariance property of the odds ratio. However, it has long been questioned whether the emphasis on relative risk combined with the rare disease assumption is relevant for public policies: see, e.g., Hsieh, Manski, and McFadden (1985) and Manski (2009) among others. We take a pragmatic approach to this debate and believe that both attributable risk and relative risk are useful for evidence-based policy-making.

References

Ackerberg, D., X. Chen, J. Hahn, and Z. Liao (2014): "Asymptotic efficiency of semiparametric two-step GMM," *Review of Economic Studies*, 81(3), 919–943.

Ai, C., and X. Chen (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71(6), 1795–1843.

——— (2012): "The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions," *Journal of Econometrics*, 170(2), 442–457.

BHATTACHARYA, J., A. M. SHAIKH, AND E. VYTLACIL (2008): "Treatment effect bounds under monotonicity assumptions: an application to Swan-Ganz catheterization," *American Economic Review: Papers and Proceedings*, 98(2), 351–56.

BHATTACHARYA, J., A. M. SHAIKH, AND E. VYTLACIL (2012): "Treatment effect bounds: An application to Swan-Ganz catheterization," *Journal of Econometrics*, 168(2), 223–243.

BRESLOW, N. E. (1996): "Statistics in epidemiology: the case-control study," *Journal of the American Statistical Association*, 91(433), 14–28.

BRESLOW, N. E., AND N. E. DAY (1980): *Statistical Methods in Cancer Research I. The Analysis of Case-Control Studies*, vol. 1. International Agency for Research on Cancer, Lyon, France.

BRESLOW, N. E., J. M. ROBINS, AND J. A. WELLNER (2000): "On the semiparametric efficiency of logistic regression under case-control sampling," *Bernoulli*, 6(3), 447–455.

CARVALHO, L. S., AND R. R. SOARES (2016): "Living on the edge: Youth entry, career and exit in drug-selling gangs," *Journal of Economic Behavior & Organization*, 121, 77–98.

CHEN, H. (2007): "A semiparametric odds ratio model for measuring association," *Biometrics*, 63(2), 413–421.

CHEN, K. (2001): "Parametric models for response-biased sampling," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4), 775–789.

CHEN, X. (2007): "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," vol. 6 of *Handbook of Econometrics*, pp. 5549–5632. Elsevier.

CHEN, X., O. LINTON, AND I. VAN KEILEGOM (2003): "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 71(5), 1591–1608.

CORNFIELD, J. (1951): "A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix," *Journal of the National Cancer Institute*, 11(6), 1269–1275.

COSSLETT, S. R. (1981): "Maximum Likelihood Estimator for Choice-Based Samples," *Econometrica*, 49(5), 1289–1316.

CURRIE, J., AND M. NEIDELL (2005): "Air pollution and infant health: what can we learn from California's recent experience?," *Quarterly Journal of Economics*, 120(3), 1003–1030.

DELAVANDE, A., AND B. ZAFAR (2019): "University Choice: The Role of Expected Earnings, Nonpecuniary Outcomes, and Financial Constraints," *Journal of Political Economy*, 127(5), 2343–2393.

DOMOWITZ, I., AND R. L. SARTAIN (1999): "Determinants of the Consumer Bankruptcy Decision," *Journal of Finance*, 54(1), 403–420.

EFRON, B. (1982): *The jackknife, the bootstrap and other resampling plans*. Society for Industrial and Applied Mathematics (SIAM).

EPPLE, D., R. E. ROMANO, AND M. URQUIOLA (2017): "School Vouchers: A Survey of the Economics Literature," *Journal of Economic Literature*, 55(2), 441–92.

HAHN, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66(2), 315–331.

HAUSMAN, J. A., AND D. A. WISE (1981): "Stratification on endogenous variables and estimation: The Gary income maintenance experiment," *Structural analysis of discrete data with econometric applications*, pp. 365–391.

HECKMAN, J. J., AND R. ROBB (1985): "Alternative methods for evaluating the impact of interventions: An overview," *Journal of econometrics*, 30(1-2), 239–267.

HECKMAN, J. J., AND P. E. TODD (2009): "A note on adapting propensity score matching and selection models to choice based samples," *Econometrics Journal*,

12(s1), S230–S234.

HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

HOLLAND, P. W., AND D. B. RUBIN (1988): "Causal Inference in Retrospective Studies," *Evaluation Review*, 12(3), 203–231.

HOROWITZ, J. L., AND C. F. MANSKI (1995): "Identification and robustness with contaminated and corrupted data," *Econometrica*, pp. 281–302.

——— (1997): "What can be learned about population parameters when the date are contaminated," *Handbook of Statistics*, 15, 439–466.

HSIEH, D. A., C. F. MANSKI, AND D. MCFADDEN (1985): "Estimation of Response Probabilities from Augmented Retrospective Observations," *Journal of the American Statistical Association*, 80(391), 651–662.

HU, Z., AND J. QIN (2018): "Generalizability of causal inference in observational studies under retrospective convenience sampling," *Statistics in Medicine*, 37(19), 2874–2883.

IMBENS, G. W. (1992): "An Efficient Method of Moments Estimator for Discrete Choice Models With Choice-Based Sampling," *Econometrica*, 60(5), 1187–1214.

IMBENS, G. W., AND T. LANCASTER (1996): "Efficient estimation and stratified sampling," *Journal of Econometrics*, 74(2), 289–318.

JUN, S. J., AND S. LEE (2019): "Identifying the effect of persuasion," arXiv:1812.02276 [econ.EM] https://arxiv.org/abs/1812.02276.

KALBFLEISCH, J., AND J. LAWLESS (1988): "Estimation of Reliability in Field–Performance Studies," *Technometrics*, 30(4), 365–378.

KENNEDY, E. H., A. SJÖLANDER, AND D. SMALL (2015): "Semiparametric causal inference in matched cohort studies," *Biometrika*, 102(3), 739–746.

KIM, W., K. KWON, S. KWON, AND S. LEE (2018): "The identification power of smoothness assumptions in models with counterfactual outcomes," *Quantitative Economics*, 9(2), 617–642.

KREIDER, B., J. V. PEPPER, C. GUNDERSEN, AND D. JOLLIFFE (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(499), 958–975.

LANCASTER, T., AND G. IMBENS (1996): "Case-control studies with contaminated controls," *Journal of Econometrics*, 71(1-2), 145–160.

MACHADO, C., A. M. SHAIKH, AND E. J. VYTLACIL (2019): "Instrumental variables and the sign of the average treatment effect," *Journal of Econometrics*, 212(2), 522–555.

MACLEOD, W. B., AND M. URQUIOLA (2019): "Is Education Consumption or Investment? Implications for School Competition," *Annual Review of Economics*, 11(1), 563–589.

MANSKI, C. F. (1995): *Identification problems in the social sciences*. Harvard University Press.

——— (1997): "Monotone Treatment Response," *Econometrica*, 65(6), 1311–1334.

——— (2003): *Partial identification of probability distributions*. Springer-Verlag.

——— (2009): *Identification for prediction and decision*. Harvard University Press.

MANSKI, C. F., AND S. R. LERMAN (1977): "The Estimation of Choice Probabilities from Choice Based Samples," *Econometrica*, 45(8), 1977–1988.

MANSKI, C. F., AND D. MCFADDEN (1981): "Alternative estimators and sample designs for discrete choice analysis," in *Structural analysis of discrete data with econometric applications*, ed. by C. F. Manski, and D. McFadden, vol. 2, pp. 51–111. MIT Press, Cambridge, MA.

MANSKI, C. F., AND J. V. PEPPER (2000): "Monotone instrumental variables: With an application to the returns to schooling," *Econometrica*, 68(4), 997–1010.

MÅNSSON, R., M. M. JOFFE, W. SUN, AND S. HENNESSY (2007): "On the estimation and use of propensity scores in case-control and case-cohort studies," *American journal of epidemiology*, 166(3), 332–339.

MCFADDEN, D. (2015): "Observational Studies: Outcome-Based Sampling," in *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, ed. by J. D. Wright, pp. 103–106. Elsevier, Oxford.

MOLINARI, F. (2020): "Microeconometrics with Partial Identification," arXiv:2004.11751 [econ.EM] https://arxiv.org/abs/2004.11751.

NEWEY, W. K. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5(2), 99–135.

——— (1994): "The asymptotic variance of semiparametric estimators," *Econometrica*, 62(6), 1349–1382.

OKUMURA, T., AND E. USUI (2014): "Concave-monotone treatment response and monotone treatment selection: With an application to the returns to schooling," *Quantitative Economics*, 5(1), 175–194.

ROSE, S. (2011): "Causal inference for case-control studies," Ph.D. thesis, Biostatistics of the UC-Berkeley.

ROSENFELD, B. (2017): "Reevaluating the Middle-Class Protest Paradigm: A Case-Control Study of Democratic Protest Coalitions in Russia," *American Political Science Review*, 111(4), 637–652.

RUGGLES, S., S. FLOOD, R. GOEKEN, J. GROVER, E. MEYER, J. PACAS, AND M. SOBEK (2019): "IPUMS USA: Version 9.0 [dataset]," https://doi.org/10.18128/D010.V9.0.

TAMER, E. (2010): "Partial identification in econometrics," *Annual Review of Economics*, 2(1), 167–195.

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(11), 1314–1316.

TCHETGEN TCHETGEN, E. J., J. M. ROBINS, AND A. ROTNITZKY (2010): "On doubly robust estimation in a semiparametric odds ratio model," *Biometrika*, 97(1), 171–180.

VAN DER LAAN, M. J., AND S. ROSE (2011): *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

XIE, J., Y. LIN, X. YAN, AND N. TANG (2020): "Category-Adaptive Variable Screening for Ultra-High Dimensional Heterogeneous Categorical Data," *Journal of the American Statistical Association*, 115(530), 747–760.

XIE, Y., AND C. F. MANSKI (1989): "The logit model and response-based samples," *Sociological Methods & Research*, 17(3), 283–302.

ZHANG, Z., Z. HU, AND C. LIU (2019): "Estimating the Population Average Treatment Effect in Observational Studies with Choice-Based Sampling," *The international journal of biostatistics*, 15(1).

ZHOU, J., A. H. HERRING, A. BHATTACHARYA, A. F. OLSHAN, D. B. DUNSON, AND N. B. D. P. STUDY (2016): "Nonparametric Bayes modeling for case control studies with many predictors," *Biometrics*, 72(1), 184–192.

APPENDICES

The appendices include aggregation of the odds ratio without taking the logarithm, a small Monte Carlo experiment, computational algorithms that are omitted from the main text, and all the proofs.

APPENDIX A. AVERAGING WITHOUT TAKING THE LOGARITHM

In the main text we followed the convention of using the *logarithm* of the odds ratio, and our key estimand was an aggregated version of the *logarithm* of the odds ratio, i.e. $\beta(y) = \mathbb{E}[\log\{\mathrm{OR}(X)\}|Y = y]$ for $y = 0, 1$. As a result, the central causal parameter under discussion was the logarithm of relative risk.

Alternatively, one may want to proceed without taking the logarithm, which does not change the substance of our results. Specifically, in this case, we are led to consider

$$\overline{\zeta} := \mathbb{E}\big[\theta(X^*)\big], \;\; \zeta(y) := \int_{\mathcal{X}} \theta(x) dF_{X|Y}(x|y), \;\; \kappa(y) := \int_{\mathcal{X}} \mathrm{OR}(x) dF_{X|Y}(x|y)$$

for $y = 0, 1$. Again, if the MTR and MTS conditions are satisfied, then we have

$$1 \leq \zeta(y) \leq \kappa(y) \tag{A.1}$$

under both designs 1 and 2, where the inequalities are sharp. Efficient estimation of $\kappa(y)$ can be explored exactly in the same way as in section 5.2. Below we present the formula of the efficient influence function, which is an analog of theorem 4.

Theorem A.1. *Suppose that assumptions A, G and H hold and that we have a sample by Bernoulli sampling. Then, for $y = 0, 1$, $\kappa(y)$ is pathwise differentiable and its pathwise derivative is given by*

$$K_y(Y, T, X) = \frac{Y^y(1-Y)^{1-y}}{h_0^y(1-h_0)^{1-y}}\Big\{\mathrm{OR}(X) - \kappa(y)\Big\} - \mathrm{OR}(X)\frac{\Delta_0(Y, T, X)}{(1-h_0)w(X)^y} + \mathrm{OR}(X)\frac{w(X)^{1-y}\Delta_1(Y, T, X)}{h_0}.$$

Further, K_y is an element of the tangent space, and therefore, the semiparametric efficiency bound for $\kappa(y)$ is given by $\mathbb{E}\{K_y^2(Y,T,X)\}$.

We can construct efficient estimators of $\kappa(y)$ and carry out causal inference on $\overline{\zeta}$ by methods identical to those used in section 5.2. We do not repeat all the details for brevity.

In general we have the relationship

$$\bar{\vartheta} := \mathbb{E}\{\log \mathrm{OR}(X^*)\} \leq \log\big[\mathbb{E}\{\mathrm{OR}(X^*)\}\big]$$

by Jensen's inequality. Therefore, an average of the log odds ratio is less likely to be affected unduly by outliers than that of the odds ratio itself. This seems to be another merit in using the logarithm of the odds ratio in addition to the usual advantage that it corresponds to the coefficients of the treatment variable when a parametric logistic model is used.

Appendix B. Monte Carlo Experiments

In this section we report the results of a small Monte Carlo experiment. A case-control sample is generated from

$$X \mid Y = y \sim \mathbb{N}(\mu^{(y)}, \Sigma^{(y)}),$$
$$\mathbb{P}(T = 1 \mid X = x, Y = y) = G(\alpha_0^{(y)} + X^\intercal \alpha_1^{(y)}),$$

where $G(u) = \exp(u)/\{1+\exp(u)\}$, $\alpha_0^{(y)}$, $\alpha_1^{(y)}$, $\mu^{(y)}$ and $\Sigma^{(y)}$ are parameters that may depend on $y = 0, 1$. In simulations we focus on estimating $\beta(y)$ that can now be expressed as

$$\beta(y) = (\alpha_0^{(1)} - \alpha_0^{(0)}) + \mathbb{E}(X|Y=y)^\intercal(\alpha_1^{(1)} - \alpha_1^{(0)}).$$

With the dimension of X equal to $d_x = 5$, the parameter values are specified as follows: $\mu^{(1)} = (1,\ldots,1)^\intercal$, $\mu^{(0)} = (0,\ldots,0)^\intercal$, and $\Sigma^{(y)} = \Sigma$ for $y = 0, 1$, where the (j,k) element of Σ is $\Sigma_{j,k} = \rho^{|j-k|}$ and $\rho = 0.5$; $\alpha_0^{(1)} = 0.5$, $\alpha_1^{(1)} = (1,1,0,0,0)^\intercal$, $\alpha_0^{(0)} = 0$, $\alpha_1^{(0)} = (0,0,1,1,0)^\intercal$. In this design we have $\beta(1) = \beta(0) = 0.5$.

In each Monte Carlo replication, we simulate 1,000 observations separately for both $Y = 0$ and $Y = 1$ samples (that is, the total sample size is 2,000 and $\hat{h} = 0.5$). There were 1,000 Monte Carlo replications.

TABLE A.1. Results of Monte Carlo Experiments

	$\beta(1)$		$\beta(0)$	
	parametric	sieve	parametric	sieve
Mean Bias	0.011	0.070	0.005	0.046
Median Bias	0.012	0.086	-0.001	0.042
RMSE	0.057	0.167	0.033	0.067
Mean AD	0.191	0.330	0.145	0.206
Median AD	0.160	0.283	0.119	0.173
Cov. Prob.	0.944	0.962	0.952	0.962

Note: RMSE stands for the root mean squared error and AD refers to absolute deviation. Cov. Prob. is the coverage probability of the one-sided 95% confidence interval. The results are based on 1,000 Monte Carlo repetitions.

In the Monte Carlo experiment, two estimators are considered: (i) a retrospective parametric logistic estimator that uses X as covariates and (ii) a retrospective sieve logistic estimator that uses the linear, quadratic and interaction terms of X as covariates (that is, $2d_x + d_x(d_x - 1)/2 = 20$ covariates all together). Table A.1 summarizes the results of the Monte Carlo experiments. Not surprisingly, the parametric estimator performs better for both $\beta(1)$ and $\beta(0)$. It shows almost no bias and small root mean squared errors and absolute deviations. Its coverage probability is close to the nominal 95%. The sieve estimator exhibits some positive biases but its performance is overall satisfactory.

APPENDIX C. DETAILS OF CAUSAL INFERENCE ON ATTRIBUTABLE RISK

This part of the appendix provides computational details of the bootstrap-based causal inference procedure on attributable risk we discussed in section 7. We focus on design 1 and below we present the algorithm.

Algorithm 2: Causal Inference on Attributable Risk Using Case-Control Samples

Input: $\{(Y_i, T_i, X_i) : i = 1, \ldots, n\}$, the number ($B$) of bootstrap replications, the coverage probability $(1 - \alpha)$ of the confidence interval, the upper bound ($\bar{p}$) on the unknown true case probability

Output: point estimates $\widehat{\mathrm{UB}}_{\mathrm{AR}}(p)$ of the upper bounds on causal attributable risk and the upper end points of the one-sided pointwise bootstrap confidence intervals $q^*_{(1-\alpha)}(p)$ for $p \in [0, \bar{p}]$

1 Construct a grid $\mathcal{P} := \{p_0, p_1, \ldots, p_J\}$ of $[0, \bar{p}]$, where $0 = p_0 < p_1 < \cdots < p_J = \bar{p}$;

2 For each $p \in \mathcal{P}$, evaluate sample analogs $\widehat{\mathrm{UB}}_{\mathrm{AR}}(p)$ of $\mathrm{UB}_{\mathrm{AR}}(p) := (1 - p)\beta_{\mathrm{AR}}(p, 0) + p\beta_{\mathrm{AR}}(p, 1)$; in this step we need to compute retrospective estimates of $\Pi(t \mid y, X_i)$ for $t = 0, 1,\ y = 0, 1$ as well as prospective estimates of $\mathbb{P}(Y = 1 \mid X_i)$ as the definition of $\beta_{\mathrm{AR}}(p, y)$ in equation (22) shows;

3 For each bootstrap replication $b = 1, \ldots, B$, generate a bootstrap sample $\{(Y_i^{*,b}, T_i^{*,b}, X_i^{*,b}) : i = 1, \ldots, n\}$ and obtain a bootstrap estimate $\widehat{\mathrm{UB}}^{*,b}_{\mathrm{AR}}(p)$ for each $p \in \mathcal{P}$;

4 For each $p \in \mathcal{P}$, compute

$$\mu^*(p) := \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left\{ \widehat{\mathrm{UB}}^{*,b}_{\mathrm{AR}}(p) \leq \widehat{\mathrm{UB}}_{\mathrm{AR}}(p) \right\},$$

where $\mathbb{1}\{\cdot\}$ is the usual indicator function;

5 For each $p \in \mathcal{P}$, obtain $\nu^*(p) := \Phi\left[\Phi^{-1}(1 - \alpha) + 2\Phi^{-1}\{\mu^*(p)\}\right]$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable;

6 For each $p \in \mathcal{P}$, compute the $\nu^*(p)$ empirical quantile of $\widehat{\mathrm{UB}}^{*,b}_{\mathrm{AR}}(p)$, say $q^*_{(1-\alpha)}(p)$, as the $(1 - \alpha)$ one-sided pointwise bootstrap confidence interval

It is straightforward to modify algorithm 2 for case-population sampling. It suffices to replace the upper bound estimates $\widehat{\mathrm{UB}}_{\mathrm{AR}}(p)$ with those of $p\xi_{\mathrm{CP}}$, where ξ_{CP} is defined in equation (23). The resulting bootstrap confidence intervals will be a uniform confidence band over $p \in [0, \bar{p}]$.

Appendix D. Auxiliary results

Lemma A.1. *We have* $r(x, p_0) = \mathbb{P}(Y^* = 1 \mid X^* = x)$.

Proof. We focus on design 1; design 2 is similar but simpler. By the Bayes rule and the sampling design,

$$\begin{aligned}\mathbb{P}(Y^* = 1 \mid X^* = x) &= \frac{p_0 f_{X^*|Y^*}(x \mid 1)}{p_0 f_{X^*|Y^*}(x \mid 1) + (1-p_0) f_{X^*|Y^*}(x \mid 0)} \\ &= \frac{p_0 f_{X|Y}(x \mid 1)}{p_0 f_{X|Y}(x \mid 1) + (1-p_0) f_{X|Y}(x \mid 0)}. \end{aligned} \tag{A.2}$$

Here, by the Bayes rule again, for $y = 0, 1$,

$$f_{X|Y}(x|y) = \frac{f_X(x)\mathbb{P}(Y = y \mid X = x)}{\mathbb{P}(Y = y)}. \tag{A.3}$$

Combining equations (A.2) and (A.3) yields the result. □

Lemma A.2. *For $t \in \{0, 1\}$, we have*

$$\mathbb{P}(Y^* = 1 \mid T^* = t, X^* = x) = \frac{r(x, p_0)\Pi(t \mid 1, x)}{\Pi(t \mid 0, x) + r(x, p)\{\Pi(t \mid 1, x) - \Pi(t \mid 0, x)\}},$$

where $p = p_0$ under design 1 and $p = 0$ under design 2.

Proof. First, consider design 1. By the Bayes rule and lemma A.1,

$$\mathbb{P}(Y^* = 1 \mid T^* = t, X^* = x) = \frac{r(x, p_0)\Pi(t \mid 1, x)}{r(x, p_0)\Pi(t \mid 1, x) + \{1 - r(x, p_0)\}\Pi(t \mid 0, x)}.$$

Under design 2, the Bayes rule yields,

$$\mathbb{P}(Y^* = 1 \mid T^* = t, X^* = x) = \frac{r(x, p_0)\Pi(t \mid 1, x)}{\Pi(t \mid 0, x)}.$$

Note that $r(x, 0) = 0$ under design 2. □

Lemma A.3. *We have*

$$\frac{\mathbb{P}(Y^* = 1 \mid T^* = 1, X^* = x)}{\mathbb{P}(Y^* = 1 \mid T^* = 0, X^* = x)} = \begin{cases} \Gamma(x, p_0) & \text{under design 1,} \\ \Gamma(x, 0) & \text{under design 2.} \end{cases}$$

Proof. It directly follows from the definitions of Γ and r, and lemma A.2. □

Lemma A.4. *Suppose that assumption E holds. Then, for $t = 0, 1$ and for all $x \in \mathcal{X}$,*

$$(-1)^t \big[\mathbb{P}\{Y^*(t) = 1 \mid X^* = x\} - \mathbb{P}(Y^* = 1 \mid X^* = x)\big] \leq 0,$$

where the bounds are sharp.

Proof. Since the two inequalities are similar, we focus on the case of $t = 1$. In this case, the claimed inequality follows from

$$\begin{aligned}\mathbb{P}\{Y^*(1) = 1, T^* = 1 \mid X^* = x\} + \mathbb{P}\{Y^*(1) = 1, T^* = 0 \mid X^* = x\} \\ \geq \mathbb{P}\{Y^*(1) = 1, T^* = 1 \mid X^* = x\} + \mathbb{P}\{Y^*(0) = 1, T^* = 0 \mid X^* = x\}.\end{aligned}$$

For sharpness, we know from assumption E that

$$\begin{aligned}\mathbb{P}\{Y^*(1) = 1, T^* = 0 \mid X^* = x\} - \mathbb{P}\{Y^*(0) = 1, T^* = 0 \mid X^* = x\} \\ = \mathbb{P}\{Y^*(1) = 1, Y^*(0) = 0, T^* = 0 \mid X^* = x\},\end{aligned}$$

where the right–hand side is unrestricted between 0 and 1. □

Lemma A.5. *Suppose that assumption F holds. Then, for $t = 0, 1$ and for all $x \in \mathcal{X}$,*

$$(-1)^t \big[\mathbb{P}\{Y^*(t) = 1 \mid X^* = x\} - \mathbb{P}(Y^* = 1 \mid T^* = t, X^* = x)\big] \geq 0,$$

where the bounds are sharp.[14]

Proof. Since the two inequalities are similar, we focus on the case of $t = 1$. First,

$$\begin{aligned}\mathbb{P}\{Y^*(1) = 1 \mid X^* = x\} = \mathbb{P}(Y^* = 1 \mid T^* = 1, X^* = x)\mathbb{P}(T^* = 1 \mid X^* = x) \\ + \mathbb{P}\{Y^*(1) = 1 \mid T^* = 0, X^* = x\}\mathbb{P}(T^* = 0 \mid X^* = x), \quad \text{(A.4)}\end{aligned}$$

where we note from assumption F that there exists some $C_x \in [0, 1]$ such that

$$\mathbb{P}(Y^* = 1 \mid T^* = 1, X^* = x) = \mathbb{P}\{Y^*(1) = 1 \mid T^* = 0, X^* = x\} + C_x. \quad \text{(A.5)}$$

[14] Furthermore, the proof shows that if $0 < \mathbb{P}(T^* = 1 | X^* = x) < 1$, these inequalities hold with equality if and only if assumption F is satisfied with equality.

Combining equations (A.4) and (A.5) yields the first inequality in the lemma statement. Therefore,

$$\begin{aligned}\mathbb{P}\{Y^*(1) = 1 \mid X^* = x\} &= \mathbb{P}(Y^* = 1 \mid T^* = 1, X^* = x) - C_x \cdot \mathbb{P}(T^* = 0 \mid X^* = x) \\ &\leq \mathbb{P}(Y^* = 1 \mid T^* = 1, X^* = x). \quad \text{(A.6)}\end{aligned}$$

Sharpness follows from the fact that C_x is not restricted except that it is between 0 and 1. Also, if $\mathbb{P}(T^* = 0|X^* = x) > 0$, then the last inequality in equation (A.6) holds with equality if and only if $C_x = 0$. □

Lemma A.6. *Suppose that assumptions E and F hold. Then, under design 1, for all $x \in \mathcal{X}$, we have $\Gamma(x, p_0) \leq \Gamma(x, 0)$.*

Proof. Recall from lemma A.3 that under design 1

$$\Gamma(x, p_0) = \frac{\mathbb{P}(Y^* = 1 \mid T^* = 1, X^* = x)}{\mathbb{P}(Y^* = 1 \mid T^* = 0, X^* = x)}. \quad \text{(A.7)}$$

Therefore, we can write

$$\Gamma(x, 0) = \Gamma(x, p_0) \frac{\mathbb{P}(Y^* = 0 \mid T^* = 0, X^* = x)}{\mathbb{P}(Y^* = 0 \mid T^* = 1, X^* = x)}. \quad \text{(A.8)}$$

So, it suffices to show

$$\mathbb{P}(Y^* = 1 \mid T^* = 1, X^* = x) \geq \mathbb{P}(Y^* = 1 \mid T^* = 0, X^* = x), \quad \text{(A.9)}$$

but this directly follows from assumptions E and F. □

Appendix E. Proofs of the results in the main text

Proof of theorem 1: First, consider $\theta(x)$. By assumption C,

$$\theta(x) = \frac{\mathbb{P}(Y^* = 1 \mid T^* = 1, X^* = x)}{\mathbb{P}(Y^* = 1 \mid T^* = 0, X^* = x)}, \quad \text{(A.10)}$$

to which we apply lemma A.3. □

Proof of theorem 2: It immediately follows from theorem 1. □

Proof of theorem 3: The sharp lower bounds of $\theta(x)$ under assumption E follows from lemma A.4. Further, from lemma A.5, we know that

$$\theta(x) \leq \frac{\mathbb{P}(Y^* = 1 \mid T^* = 1, X^* = x)}{\mathbb{P}(Y^* = 1 \mid T^* = 0, X^* = x)}, \tag{A.11}$$

where the bounds are sharp under assumption F. So, apply lemmas A.3 and A.6. Sharpness follows from continuity. □

Proof of Lemma 1: Let γ be the parameter denoting regular parametric submodels, where the true value will be denoted by γ_0. Then, by using the likelihood function in equation (11), the score evaluated at γ_0 is equal to

$$\begin{aligned}(1-Y)\Big[S_{X|Y}(X|0) + \frac{\{T - \mathbb{P}(T=1|X,Y=0)\}\partial_\gamma \mathbb{P}(T=1|X,Y=0;\gamma_0)}{\mathbb{P}(T=1|X,Y=0)\{1-\mathbb{P}(T=1|X,Y=0)\}}\Big] \\ + Y\Big[S_{X|Y}(X|1) + \frac{\{T - \mathbb{P}(T=1|X,Y=1)\}\partial_\gamma \mathbb{P}(T=1|X,Y=1;\gamma_0)}{\mathbb{P}(T=1|X,Y=1)\{1-\mathbb{P}(T=1|X,Y=1)\}}\Big],\end{aligned} \tag{A.12}$$

where $S_{X|Y}(x|y) = \partial_\gamma \log f_{X|Y}(x|y;\gamma_0)$ is restricted only by $\mathbb{E}\{S_{X|Y}(X|y)|Y=y\} = 0$, while the derivatives $\partial_\gamma \mathbb{P}(T=1|X,Y=y,\gamma_0)$ are unrestricted. □

Proof of theorem 4: For brevity, we focus on $\beta(0)$ and let $\beta = \beta(0)$. Proof for $\beta(1)$ is analogous. Let $p_0(x) = \mathbb{P}(T=1|X=x,Y=0)$ and $p_1(x) = \mathbb{P}(T=1|X=x,Y=1)$. Note that

$$\beta(\gamma) = \int_{\mathcal{X}} \log \Big[\underbrace{\frac{p_1(x;\gamma)}{1-p_1(x;\gamma)} \cdot \frac{1-p_0(x;\gamma)}{p_0(x;\gamma)}}_{:=\mathrm{OR}(x;\gamma)}\Big] f_{X|Y}(x|0;\gamma)dx, \tag{A.13}$$

where γ represents regular parametric submodels such that γ_0 is the truth. Then,

$$\begin{aligned}\partial_\gamma \mathrm{OR}(x;\gamma_0) &= \partial_\gamma p_1(x;\gamma_0)\frac{\{1-p_0(x)\}}{p_0(x)\{1-p_1(x)\}^2} - \partial_\gamma p_0(x;\gamma_0)\frac{p_1(x)}{p_0^2(x)\{1-p_1(x)\}} \\ &= \frac{\partial_\gamma p_1(x;\gamma_0)}{p_1(x)\{1-p_1(x)\}}\mathrm{OR}(x) - \frac{\partial_\gamma p_0(x;\gamma_0)}{p_0(x)\{1-p_0(x)\}}\mathrm{OR}(x).\end{aligned} \tag{A.14}$$

Therefore,

$$\partial_\gamma \beta(\gamma_0) = \int \Big[\frac{\partial_\gamma \mathrm{OR}(x;\gamma_0)}{\mathrm{OR}(x)} + \log\{\mathrm{OR}(x)\} S_{X|Y}(x|0) \Big] f_{X|Y}(x|0) dx$$
$$= \int \Big[\frac{\partial_\gamma p_1(x;\gamma_0)}{p_1(x)\{1-p_1(x)\}} - \frac{\partial_\gamma p_0(x;\gamma_0)}{p_0(x)\{1-p_0(x)\}} + \log\{\mathrm{OR}(x)\} S_{X|Y}(x|0) \Big] f_{X|Y}(x|0) dx. \tag{A.15}$$

Now, we only need to verify the equality between $\mathbb{E}\{F_0(Y,T,X)S(Y,T,X)\}$ and

$$\int \Big[\underbrace{\frac{\partial_\gamma p_1(x;\gamma_0)}{p_1(x)\{1-p_1(x)\}}}_{:=A_1(x)} - \underbrace{\frac{\partial_\gamma p_0(x;\gamma_0)}{p_0(x)\{1-p_0(x)\}}}_{:=A_0(x)} + \log\{\mathrm{OR}(x)\} S_{X|Y}(x|0) \Big] f_{X|Y}(x|0) dx, \tag{A.16}$$

where $F_0(Y,T,X)$ and $S(Y,T,X)$ are given in the theorem statement and equation (A.12), respectively: i.e.

$$S(Y,T,X)$$
$$= (1-Y)\Big[S_{X|Y}(X|0) + \{T - p_0(X)\} A_0(X) \Big] + Y \Big[S_{X|Y}(X|1) + \{T - p_1(X)\} A_1(X) \Big],$$
$$F_0(Y,T,X)$$
$$= \frac{1-Y}{1-h_0} \left[\log \mathrm{OR}(X) - \beta - \frac{\{T - p_0(X)\}}{p_0(X)\{1-p_0(X)\}} \right] + \frac{Y}{h_0} \frac{f_{X|Y}(X|0)}{f_{X|Y}(X|1)} \frac{\{T - p_1(X)\}}{p_1(X)\{1-p_1(X)\}}.$$

Note that $F_0(Y,T,X)S(Y,T,X)$ is equal to

$$\frac{1-Y}{1-h_0} \left[\log \mathrm{OR}(X) - \beta - \frac{\{T - p_0(X)\}}{p_0(X)\{1-p_0(X)\}} \right] \Big[S_{X|Y}(X|0) + \{T - p_0(X)\} A_0(X) \Big]$$
$$+ \frac{Y}{h_0} \frac{f_{X|Y}(X|0)}{f_{X|Y}(X|1)} \left[\frac{\{T - p_1(X)\}}{p_1(X)\{1-p_1(X)\}} \right] \Big[S_{X|Y}(X|1) + \{T - p_1(X)\} A_1(X) \Big].$$

Here, taking expectations directly shows that $\mathbb{E}\big\{F_0(Y,T,X)S(Y,T,X)\big\}$ is equal to

$$\mathbb{E}\big\{ \log\{\mathrm{OR}(X)\} S_{X|Y}(X|0) - A_0(X) | Y = 0 \big\} + \mathbb{E}\left\{ \frac{f_{X|Y}(X|0)}{f_{X|Y}(X|1)} A_1(X) \middle| Y = 1 \right\},$$

which is equal to the expression in equation (A.16) since

$$\mathbb{E}\left\{\frac{f_{X|Y}(X|0)}{f_{X|Y}(X|1)}A_1(X)\middle|Y=1\right\} = \mathbb{E}\big\{A_1(X)|Y=0\big\}.$$

Finally, it follows from lemma 1 that F_0 is an element of the tangent space. □

Proof of theorem 5: Under strong ignorability, we have

$$\theta_{\mathrm{AR}}(x) = \mathbb{P}(Y^*=1 \mid T^*=1, X^*=x) - \mathbb{P}(Y^*=1 \mid T^*=0, X^*=x).$$

Therefore, the statement follows from lemma A.2 and the definition of Γ_{AR}. □

Proof of theorem 6: It immediately follows from theorem 5. □

Proof of theorem 7: The lower bound trivially follows by the same argument as in theorem 3. For the upper bound, lemma A.5 shows that

$$\theta_{\mathrm{AR}}(x) \leq \mathbb{P}(Y^*=1 \mid T^*=1, X^*=x) - \mathbb{P}(Y^*=1 \mid T^*=0, X^*=x),$$

where the inequality is sharp under assumption F. Then, by lemma A.2 and the definition of Γ_{AR}, the right–hand side is equal to $r(x,p_0)\Gamma_{\mathrm{AR}}(x,p_0)$ and $r(x,p_0)\Gamma_{\mathrm{AR}}(x,0)$ under designs 1 and 2, respectively. □

Proof of theorem A.1: It follows from the same arguments as in the proof of theorem 4. We omit the details. □