9-9-2009

Estimating effects by combining instrumental variables with case-control designs: the role of principal stratification

Russell T. Shinohara
Johns Hopkins University, taki.shinohara@gmail.com

Constantine E. Frangakis
Johns Hopkins University, cfrangak@jhsph.edu

Elizabeth Platz
Johns Hopkins University

Konstantinos Tsilidis
Oxford University

Johns Hopkins University, Dept. of Biostatistics Working Papers

Year 2009 Paper 198

Estimating effects by combining instrumental variables with case-control designs: the role of principal stratification

Russell T. Shinohara [1], Constantine E. Frangakis [1],
Elizabeth A. Platz [2], and Konstantinos K. Tsilidis [2,3]

September 17, 2009

Summary. The instrumental variable framework is commonly used in the estimation of causal effects from cohort samples. In the case of more efficient designs such as the case-control study, however, the combination of the instrumental variable and complex sampling designs requires new methodological consideration. As the prevalence of Mendelian randomization studies is increasing and the cost of genotyping and gene expression data can be high, the analysis of data gathered from more cost-effective sampling designs is of prime interest. We show that the standard instrumental variable analysis is not applicable to the case-control design and can lead to erroneous estimation and inference. We also propose a method based on principal stratification for the analysis of data arising from the combination of case-control sampling and instrumental variable design and illustrate it with a study in oncology.

Key words: Case-Control, Instrumental Variables, Mendelian randomization, Principal Stratification, Study Design

[1] Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA
[2] Department of Epidemiology, Johns Hopkins University, Baltimore, MD 21205, USA
[3] Cancer Epidemiology Unit, University of Oxford, Richard Doll Building, Oxford, OX3 7LF, UK

1. Introduction

Many scientific studies seek to estimate the causal effect of an exposure E on an outcome Y. Although experimentation with E is the most reliable design for making causal claims, there is a plethora of situations in which it is either impossible or unfeasible. To address this, it is often useful to combine designs.

An example of such a combined design is involved in the CLUE II study (e.g., see Erlinger et al. 2004), a prospective cohort follow-up of serological risk factors for cancer and heart disease. The subjects were sampled as a large cohort from communities in Maryland between 1989 and 2000. The scientific target was the effect that chronic low-grade inflammation has on the risk of colorectal cancer. Inflammation, was measured via C-reactive protein (CRP) concentrations in archived baseline blood samples. For reasons common to many such studies, two designs were employed simultaneously.

First, the cost of measuring some variables made it infeasible to record all such information for each member of the cohort. A nested case-control design was thus employed. All 172 cases of colorectal cancer were sampled and 342 matched controls were selected based on age, sex, and date of blood draw.

Secondly, the mechanism of inflammation is not fully known. For this reason we use the genotype G of each member of the case-control study for a given SNP called rs1205 that is known to stimulate inflammatory processes (Carlson et al., 2005) as the variation in G is well understood and stems from meiosis and fertilization. Details on the genotyping and SNP selection process have been published elsewhere (Tsilidis et al., 2009). In the more general literature, such genetic variables are also known as Mendelian randomizers (see Didelez and Sheehan, 2007, or Davey Smith and Ebrahim, 2003, for an extensive review). The above design therefore combines case-control and mendelian randomization.

We show here that the methodology of standard instrumental variables is not directly applicable to estimate effects when mendelian randomization is combined with case-control designs. This is because the case control data alone do not include any information about the prevalence (or incidence) of disease in the population [1]. We demonstrate that in order to identify the causal effect of interest, the data require supplementation with the incidence of cancer. This result is important when contrasted with the case of more classical analyses involving the estimation of odds ratios of association, where such incidence information is ancillary (Prentice and Pike, 1979). In the combined case-control mendelian randomization, however, the incidence $P(Y)$ plays a crucial role.

We also show how to address this problem. First, as the prevalence information is not available from the case-control study alone, it must be obtained or estimated externally. Such information is often readily available from prior studies directly or through databases such as the Surveillance Epidemiology and End Results (SEER, 2009) Program which lists the estimated incidence of cancer in the United States. This information must be combined with the observed case-control data in order to be analyzed properly. We show that this can be accomplished through the use of principal stratification (Frangakis and Rubin, 2002), a framework which generalized the work of Angrist, Imbens, and Rubin (1996) for instrumental variables in broader settings.

The remainder of this paper is organized as follows: in the next section we define the target of interest in terms of potential outcomes and principal strata. In Section 3, we express the data arising from the combined design in terms of principal strata: first, we express the data that arise from the cohort, for which we review the framework of instrumental variables analysis of Angrist et al. (1996) for a cohort design; second, we express the data that arise from the case-control sampling from the cohort. In Section 4 we show that standard instrumental variables analysis does not correctly estimate the causal effect of interest in the combined design. We also show how one can correctly estimate the causal effect. We apply our methods to the CLUE II study. We conclude with a discussion in Section 5.

[1] Note that the term prevalence here is used as $P(Y)$, which may also represent incidence over a time period as in the colorectal cancer study.

2. Scientific Target

We wish to estimate the effect of inflammation on colorectal cancer risk, using the SNP rs1205 as an instrumental variable in the case-control design. To do this, first consider a population of units i representable by the original *cohort* sample (for example, the CLUE II cohort sample). Consider that each individual's genotype g could have been either $g = 0$ (the rs1205 genotype (CT/TT) associated with less inflammation) or $g = 1$ (the genotype (CC) associated with more inflammation). Let $E_i(g)$ be the level of inflammation that the i-th individual would experience at a time 1 if the genotype were g. Further, let $Y_i(g)$ denote the colorectal cancer status of the i-th individual at a time 2 if the genotype were g.

To formalize the causal effect of interest, it is important to consider the strata S defined by $(E_i(0), E_i(1))$, known as the principal strata (Frangakis and Rubin, 2002). Specifically, there can be individuals who express higher inflammation with either genotype, whom we call always-inflamed and denote by $S = a$. There can also be individuals who express lower inflammation with either genotype, whom we call by never-inflamed and denote by $S = n$. There can even be individuals who express high inflammation if and only if $g = 0$ (subjects known as defiers, denoted by $S = d$). Arguably most important for our goal are the individuals who would be inflamed with genotype $g = 1$ but who would not be inflamed with genotype $g = 0$; the latter individuals we call preventable and denoted by $S = p$. For preventable subjects, the comparison of cancer status between $g = 1$ versus $g = 0$ is equivalent to that comparison between high expression $E = 1$ versus low expression $E = 0$. In contrast, the always-inflamed and never-inflamed subjects, whose CRP would not be affected by genotype, carry no information to the estimation of the effect of inflammation on cancer. For this reason, we define the effect of inflammation on the risk of colorectal cancer to be the principal causal effect (Frangakis and Rubin, 2002)

$$P(Y_i(1) \mid S_i = p) \text{ versus } P(Y_i(0) \mid S_i = p) \tag{1}$$

that is, the comparison between the risk of cancer for compliers with higher inflammation genotype versus lower inflammation genotype.

The principal strata S are not always observable so the effect (1) must be estimated from data. Thus, in the following section, we examine how (1) is connected to the likelihood of the observed data that arise from the design: first from the cohort sampling from the population, and then from the case-control subsampling from the cohort sample.

3. Case-Control Design with Instrumental Variables

3.1 *Design summary*

The design is summarized in Figure 1. For each participant, the actual genotype takes its value at meiosis and fertilization at time 0 and is denoted by G_i. The actual inflammation level measured by circulating C-reactive protein concentration takes its value at a later time 1 and is denoted by $E_i(= E_i(G_i))$. At this time 1, blood is drawn from each participant and stored for possibly measuring E_i and G_i depending on later information.

Specifically, at time 2, the actual cancer status is measured and is denoted by $Y_i(= Y_i(G_i))$. The past values of G_i (unchanged from time 0) and E_i (from time 1) are then measured in the stored blood for all cases ($Y_i = 1$), and for 2 controls ($Y_i = 0$) that match each case on covariates X_i. For simplicity, assume that each of G_i and E_i are binary. The observed data therefore are $\{X_i, G_i, E_i\}$ for all cases and an $X-$ matched sample of controls.

Next, we discuss the link between the scientific target and these data.

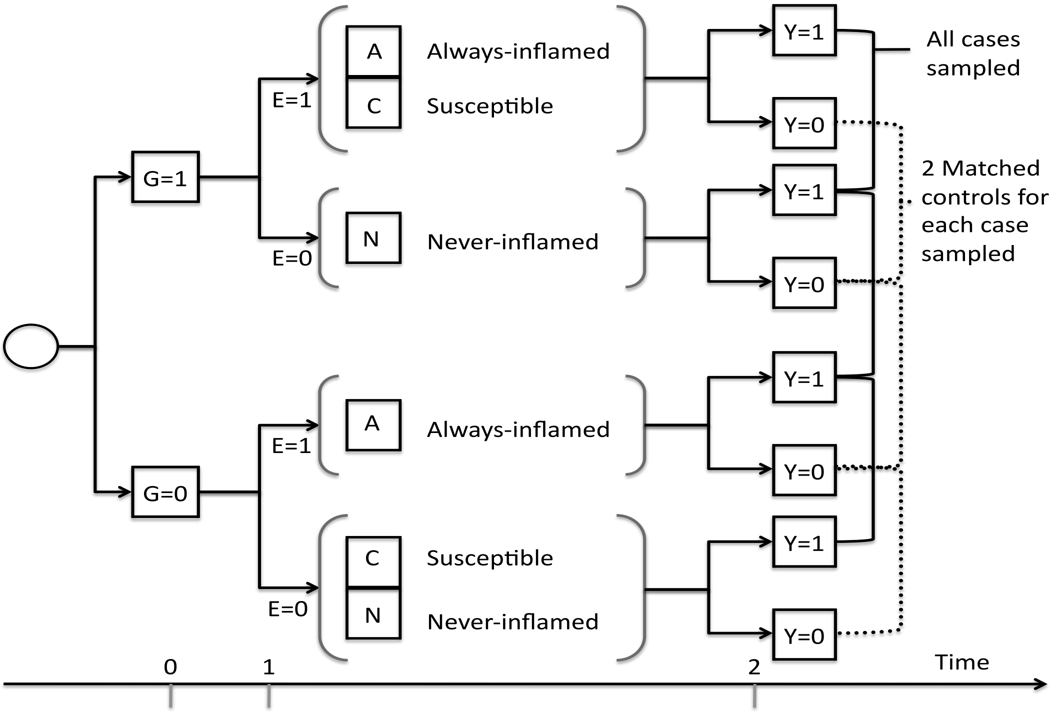

Figure 1: Summary of the design of case-control sampling in the context of instrumental variables

3.2 *Cohort Component of Design*

We consider the study cohort at time 0 to be a simple random sample of the population cohort of interest (as in Section 2), so that $(X_i, G_i, E_i(g), Y_i(g))$ are independent and identically distributed from that population. For the population cohort, we make a set of assumptions that are important for using the genotype as an instrumental variable in the sense of Angrist et al. (1996), and that are scientifically plausible in our case.

A first condition assumes that subjects with different genotypes are comparable within levels of X_i in the sense of:

Assumption 1. *Ignorability*

$$G_i \perp (Y_i(g=0), Y_i(g=1), S_i) \mid X_i$$

In the case of Mendelian randomization, ignorability is supported by the random receipt of a paternal vs. maternal allele at meiosis. Its use within covariate levels is plausible when within X there is no population admixture. It is thus important to condition on important covariates X_i. To simplify notation and with no loss of generality, until Section 4 we suppose we are within a particular level of the covariates.

A second assumption is the absence of any effect of the rs1205 genotype on colorectal cancer if the genotype has no effect on inflammation, that is,

Assumption 2. *Exclusion Restriction*

$$\text{if } E_i(g=0) = E_i(g=1) \text{ then } Y_i(g=0) = Y_i(g=1)$$

This is supported by scientific knowledge of the region of the genome in which the SNP resides. Indeed, the region is known to be directly associated with inflammatory processes and there is no evidence that it is responsible for other biological mechanism (Timpson et al., 2005).

The final assumption is that there are no individuals in the study who would suffer from increased inflammation if assigned the genotype associated with lower CRP but not if assigned the alternate genotype. That is, there are no defiers:

Assumption 3. *Monotonicity:* $E_i(g=0) \leq E_i(g=1)$

Violations of monotonicity here would be inconsistent with the established biological mechanism by which the SNP g predisposes to inflammation (as above for exclusion restriction).

For our goal of combining this design with case-control sampling it is important to summarize the implications that the above assumptions have on the cohort data, as these implications follow, for example, from Angrist et al. (1996) and Frangakis and Rubin (2002). First, any given distribution of principal strata and potential outcomes in the cohort, along with a distribution of genotypes and Assumptions 1-3, induce a distribution on the cohort data $P(G, E, Y)$.

Second, the value of Assumptions 1-3 is that this operation is invertible: any given distribution $P(G, E, Y)$ of the cohort data (e.g., as can be estimated directly) is compatible with at most one distribution for the principal strata and potential outcomes $(P(G), P(S), P(Y(g) \mid S))$. We list these mappings, given as references in the Appendix, as follows.

Definition 1. *We denote by* $M^{\text{principal}}_{\text{strata}} {}^{\rightarrow}{}^{\text{cohort}}_{\text{data}}$ *the function that maps a distribution* $(P(G), P(S), P(Y(g) \mid S))$ *to the distribution of the cohort data* $P(G, E, Y)$ *as induced by Assumptions 1-3. We denote by* $\left(M^{\text{principal}}_{\text{strata}} {}^{\rightarrow}{}^{\text{cohort}}_{\text{data}}\right)^{-1}$ *the function that maps a distribution of the cohort data* $P(G, E, Y)$ *to the compatible distribution* $(P(G), P(S), P(Y(g) \mid S))$ *of interest for the target (1).*

The observed data at time 2 (Fig. 1) then arise after taking a case-control sample from the distribution $P(G, E, Y)$ arising from the mapping $\mathrm{M}^{\text{principal}}_{\text{strata}} {}^{\rightarrow}{}^{\text{cohort}}_{\text{data}}$.

3.3 *Case-Control Component of Design*

We can treat the case-control sampling at time 2 as a sampling of a group of patients after having conditioned on the outcome Y (=1 for cases, 0 for controls) and having matched on covariates X; that is, as a sampling from the distribution $P(E, G \mid Y, X)$. For simplicity, we proceed, as in the above sections, by omitting X. It is important to formalize the relationship between the joint distribution of the observed variables in the cohort and the joint distribution induced after the case-control sampling design:

Definition 2. *Let* $M^{\text{cohort}}_{\text{data}} {}^{\rightarrow}{}^{\text{case}}_{\text{control}}$ *denote the mapping that takes the distribution of observed data in the cohort* $P(Y, E, G)$ *to the induced joint distribution resulting from case-control sampling from the cohort. Namely,*

$$P^{\text{case}}_{\text{control}}(Y, E, G) = P(Y, E, G) \frac{P^{\text{case}}_{\text{control}}(Y)}{P(Y)} \tag{2}$$

In the above, $P^{\text{case}}_{\text{control}}(Y)$ is the distribution of cases that is forced in the data after the case-control sampling. In our example of CLUE II where 2 controls were matched for each case, $P^{\text{case}}_{\text{control}}(Y = 1) = \frac{1}{3}$.

We can now represent the way in which the likelihood of the observed data, on the final end of the design, is connected to the original joint distribution of the genotype, principal strata, and potential outcomes as the mapping $\mathrm{M}^{\text{principal}\rightarrow\text{cohort}}_{\text{strata}\quad\ \text{data}}$ followed by the mapping $\mathrm{M}^{\text{cohort}\rightarrow\text{case}}_{\text{data}\quad\ \text{control}}$; namely, the composite mapping $(\mathrm{M}^{\text{cohort}\rightarrow\text{case}}_{\text{data}\quad\ \text{control}}) \circ (\mathrm{M}^{\text{principal}\rightarrow\text{cohort}}_{\text{strata}\quad\ \text{data}})$ (Figure 2).

$$\begin{tikzcd}
\left\{\begin{array}{l} P(G) \\ P(S) \\ P(Y(g)|S) \end{array}\right\} \arrow[rr, phantom, "\neq"] \arrow[d, "\mathrm{M}^{\text{principal}\rightarrow\text{cohort}}_{\text{strata}\quad\ \text{data}}"] & & \left\{\begin{array}{l} P^{(\text{case-control})}(G) \\ P^{(\text{case-control})}(S) \\ P^{(\text{case-control})}(Y(g)|S) \end{array}\right\} \\
\left\{ P(Y,E,G) \right\} \arrow[rr, "\mathrm{M}^{\text{cohort}\rightarrow\text{case}}_{\text{data}\quad\ \text{control}}"] \arrow[d, equal] & & \left\{ P^{(\text{case-control})}(Y,E,G) \right\} \arrow[u, "\left(\mathrm{M}^{\text{principal}\rightarrow\text{cohort}}_{\text{strata}\quad\ \text{data}}\right)^{-1}"'] \arrow[d, equal] \\
\left\{\begin{array}{l} P(E,G|Y) \\ P(Y) \end{array}\right\} \arrow[rr, bend right] & & \left\{\begin{array}{l} P(E,G|Y) \\ P^{(\text{case-control})}(Y) \end{array}\right\}
\end{tikzcd}$$

Figure 2: Description of the problem of case-control sampling in the context of instrumental variables.

We can now use the composite mapping to show the complication of the standard instrumental variables analysis from such design, and also show how to address the problem.

4. Methodological Implications of Design on Analysis

4.1 *Analysis of Data Arising from Combined Design*

We show that the target quantities $P(Y(g)|S)$ in (1) are not invariant to case-control sampling. Based on the mappings defined in Section 3 we have the following result, proved in the Appendix:

Result. *In general, the instrumental variable mappings* $M^{\substack{\text{principal}\\ \text{strata}} \to \substack{\text{cohort}\\ \text{data}}}$ *and* $\left(M^{\substack{\text{principal}\\ \text{strata}} \to \substack{\text{cohort}\\ \text{data}}}\right)^{-1}$ *are not invariant under sampling design. That is, when applying the function*

$$\left(M^{\substack{\text{principal}\\ \text{strata}} \to \substack{\text{cohort}\\ \text{data}}}\right)^{-1} \circ \left(M^{\substack{\text{cohort}\\ \text{data}} \to \substack{\text{case}\\ \text{control}}}\right) \circ \left(M^{\substack{\text{principal}\\ \text{strata}} \to \substack{\text{cohort}\\ \text{data}}}\right) \tag{3}$$

to a particular distribution $(P(G), P(S), P(Y(g) \mid S))$*, we obtain a different result for* $P(Y(g)|S)$*.*

This means the following: suppose we use the combined case-control and instrumental variables design to measure data (equivalent to applying the rightmost two mappings of (3) on the original problem), and suppose on those data we use the the standard instrumental variables analysis (equivalent to applying the inverse mapping in the left of (3). Then, the result for the target $P(Y(g)|S)$ is different from the true target $P(Y(g)|S)$. The reasons for this are: (i) the case-control design changes the marginal distribution of the outcome, and (ii) the structure of this problem is sufficiently rich so that the marginal distribution carries information about the causal effect (1) in light of the remaining information suppled by the conditional distribution of the genotype and inflammation $P(G, E \mid Y)$.

Table 1.
An example demonstrating the range of discrepancies between the true instrumental variables effects and their naive induced values in a case-control design.

(a) Distribution of Target Quantities in Population:

	g=0		g=1
$P(S=a)=1/3$	$P(Y=1\mid S=a)=0.8$	(=)	$P(Y=1\mid S=a)=0.8$
$P(S=s)=1/3$	$P(Y(g=0)=1\mid S=p)=0.8$		$P(Y(g=1)=1\mid S=p)=0.2$
$P(S=n)=1/3$	$P(Y=1\mid S=n)$	(=)	$P(Y=1\mid S=n)^{\dagger}$
	$P(G=0)=99\%$		$P(G=1)=1\%$

(b) True causal effects and induced values in case-control sampling:

† when $P(Y=1\mid S=n)$ is	0.2	0.6	0.8
then the true Difference is	-0.6	-0.6	-0.6
and the induced Difference is	-0.59 **(0.98)***	-0.57 **(0.95)**	**NA** (!)
the true Odds Ratio is	0.06	0.06	0.06
and the induced Odds Ratio is	0.06 **(1.00)**	0.021 **(0.33)**	**NA** (!)
the true Relative Risk is	0.25	0.25	0.25
and the induced Relative Risk is	0.20 **(0.80)**	0.050 **(0.20)**	**NA** (!)

* The ratios of induced to true values are shown in bold face font in parentheses.
(!) Note: the NAs mean that the induced distribution if one ignored the case-control design could not even be inverted (producing negative numbers for the target probabilities).

An example of the discrepancy between the true target quantities and those induced by case-control sampling is shown in Table 1. The top section of the table lists a set of fixed values for all parameters in $P(Y(g)|S), P(S), P(G))$ except $P(Y=1|S=n)$. We show that depending on the choice of $P(Y=1|S=n)$, the case-control sampling distorts the quantities of interest to varying degrees. The lower section first gives the true values of the target comparisons of interest (which are fixed among these scenarios), and is followed by three situations of induced versions of those quantities. In the first, the sampling induced targets are quite similar to the truth. The second is a situation in which a large bias is induced in the relative risk as well as the odds ratio. Finally, the choice of $P(Y = 1|S = n) = 0.8$ results in a negative value for $P(Y(1) = 1|S = c)$ and hence negative relative risk and odds ratio. This shows that when the design is ignored, one can be lead to believe that the iv assumptions are incorrect when in truth they are correct.

When the target of estimation is a more standard odds ratio between an exposure and an event, that target remains the same when changing the marginal distribution of the cases; namely, the naive odds ratio is invariant to case-control sampling (Prentice and Pike, 1979). This is not the case for instrumental variable estimands as seen in the above example.

The mappings described in Section 3 give rise to a method for estimating the causal effect arising from combination case-control instrumental variable designs. Namely,

$$\begin{aligned}&(P(G), P(S), P(Y(g) \mid S)) \\ &\qquad = \left(\mathrm{M}_{\substack{\text{principal}\\ \text{strata}} \to \substack{\text{cohort}\\ \text{data}}}\right)^{-1} \circ \left(\mathrm{M}_{\substack{\text{cohort}\\ \text{data}} \to \substack{\text{case}\\ \text{control}}}\right)^{-1} \left\{P^{\text{case}}_{\text{control}}(E, G \mid Y) \text{ and } P(Y)\right\} \qquad (4)\end{aligned}$$

Note that the above holds under the conditioning of all matching variables. In practice, we can use the following way to estimate the target (1) of the left hand side of (4):

First we obtain the incidence of cancer in the cohort, $P(Y)$; then, we take the distribution of the observed data $P^{\text{case}}_{\text{control}}(E, G \mid Y)$ in the case-control design and map it back, first to the cohort design, and then to the distribution of interest $(P(G), P(S), P(Y(g) \mid S))$.

More specifically, we first estimate the input information required in the right hand side of (4). We can estimate $P^{\text{case}}_{\text{control}}(E, G \mid Y, X)$ by estimating $P^{\text{case}}_{\text{control}}(G = 1 \mid Y, X)$ and $P^{\text{case}}_{\text{control}}(E = 1 \mid G, Y, X)$ using, for example, logistic regressions on the case-control data. Second, we can estimate $P(Y = 1|X)$ from any of a number of national databases reporting X-specific incidence rates for Y. Then, by multiplying the estimates of $P^{\text{case}}_{\text{control}}(E = 1 \mid G, Y, X)$, $P^{\text{case}}_{\text{control}}(G = 1 \mid Y, X)$ and $P(Y \mid X)$, we obtain an estimate of the distribution $P(E, G, Y \mid X)$ in the cohort (namely, we have applied the inverse of $\mathrm{M}^{\substack{\text{cohort}\\ \text{data}} \to \substack{\text{case}\\ \text{control}}}$). Finally, on the result of this calculation we can apply $\left(\mathrm{M}^{\substack{\text{principal}\\ \text{strata}} \to \substack{\text{cohort}\\ \text{data}}}\right)^{-1}$ to get a consistent estimate of $(P(G|X), P(S|X), P(Y(g) \mid S, X))$, and hence of the target causal effect (4). Estimates of the standard errors can be obtained by the delta method or by the bootstrap.

4.2 *Application to the CLUE II Study*

An example of such a study is the CLUE II cohort that has been introduced in the past sections. Data were recorded on age, sex, and date of blood draw for each participant. As the relationship between the date of blood draw and colorectal cancer can be reasonably assumed to be independent, it was omitted from the covariates X selected for the analysis. In order to model the joint distribution in the case-control study, the conditional models corresponding to $P(E|Y, X)$ and $P(G|E, Y, X)$ were fitted via logistic regression.

The estimation of the target quantities in the colorectal oncology study was conducted with and without addressing the sampling design. To obtain the adjusted estimates described in the previous section, we used the SEER incidence estimates for $P(Y \mid X)$. As an example, Table 2 shows the results for the median covariate values. As these data were part of a pilot study and the sample size is relatively small, the standard errors estimated via the delta method were large. Nevertheless, the point estimates demonstrate the possible differences that are plausible between the results of the two methods. In this case, the odds ratios between the two methods are almost the same, suggesting that there is no considerable confounding in the exogenous inflammation measurement. On the other hand, the causal effects quantified by the difference and by the relative risk are measurably different when addressing versus not addressing the combined design.

Table 2.
Results from the estimation of causal quantities of interest via unadjusted and design-adjusted procedures. The covariates age and sex were fixed at their respective medians (women of age 66).

	Estimate with adjustment	Estimate without adjustment
P(Y(0)=1|S=c)	4.28 x 10^{-4}	0.27
P(Y(1)=1|S=c)	7.76 x 10^{-4}	0.40
Difference	3.48 x 10^{-4}	0.13
Odds Ratio	1.81	1.80
Relative Risk	1.81	1.48

5. Discussion

The focus of this article is the combination of case-control and instrumental variable designs. This is of interest in itself, but is especially important as the number of Mendelian randomization studies grows. As genetic data is often prohibitively expensive to measure on a large cohort, more efficient nested case-control studies are appealing. We have shown that this combination is not amenable to the standard instrumental variables analyses, and we have provided method to address this problem. It can also be seen that the methodology is applicable when combining instrumental variables with the case-cohort design which is the most common alternative to the nested case-control.

The development of methods such as those presented in this paper have a broader implication. When covariates can be measured in a cohort with instrumental variables, a number of alternative designs is generally possible. For example, it may benefit robustness to create matches of exposed versus unexposed individuals if the exposure (rather than the instrument, or the outcome) is rare. Until now, there has been no work on such alternative designs with instrumental variables. This is, arguably, because under such designs randomization is no longer preserved in the observed data, and past frameworks of instrumental variables (e.g., with error terms on the observed data) could not easily derive the implied assumptions on the observed data. Our paper suggests there is in fact a generalizable approach to use such designs: first, we place instrumental variables assumptions (such as Assumptions 1-3) on the potential outcomes and principal strata in the ideal cohort; second, we derive the mapping from those target quantities to the measurable cohort data; and, third, we derive how our proposed additional design component (e.g., case-control matching, or exposed-unexposed matching) maps the cohort measurable data to the observed data. We can then examine what extra information (e.g., as in the marginal incidence, with the case-control design) is required in order to make the composite mapping invertible.

In conclusion, cost or efficiency considerations can suggest a case-control sampling in an instrumental variables problem. The framework of principal stratification shows that standard instrumental variables methods are not generally appropriate for such composite designs, and provides methods to better estimate the causal effects.

Appendix

The Mapping $M^{\substack{\text{principal}\\\text{strata}}\rightarrow\substack{\text{cohort}\\\text{data}}}$*:*

We denote by $\mathrm{M}^{\substack{\text{principal}\\\text{strata}}\rightarrow\substack{\text{cohort}\\\text{data}}}$ the function that maps a distribution $(P(G), P(S), P(Y(g) \mid S))$ to the distribution of the cohort data $P(G, E, Y)$. Let us denote $P(Y = 1|G = g, E = e)$ by $y^{(1)}_{g=g,e=e}$. Specifically, one may note that

$$y^{(1)}_{g=1,e=1} = \frac{s_p}{s_a + s_p} P(Y(1) = 1|S = s) + \frac{s_a}{s_a + s_p} P(Y = 1|S = a), \quad \text{and} \tag{A.1}$$

$$y^{(1)}_{g=0,e=0} = \frac{s_p}{s_n + s_p} P(Y(0) = 1|S = s) + \frac{s_n}{s_n + s_p} P(Y = 1|S = n) \tag{A.2}$$

where $s_a = P(S = a)$, $s_p = P(S = p)$, and $s_n = P(S = n)$. From Figure (1) it is clear that $y_{1,0} = P(Y = 1|S = n)$ and $y_{0,1} = P(Y = 1|S = a)$.

Similarly, $P(G, E)$ may easily be written in terms of $P(G)$ and $P(S)$. That is, denoting $P(E = 1|G = g)$ by $e^{(1)}_{g=g}$, we have that:

$$e^{(1)}_{g=1} = s_p + s_a \tag{A.3}$$

$$e^{(1)}_{g=0} = s_p + s_n \tag{A.4}$$

Hence, under Assumptions (1)-(3), there is a 1-1 map from the space of distributions $(P(G), P(S), P(Y(g) \mid S))$ to joint distributions of G, E, and Y.

The Mapping $\left(M^{\substack{\text{principal}\\ \text{strata}} \rightarrow \substack{\text{cohort}\\ \text{data}}}\right)^{-1}$*:*

The inverse mapping $\left(\mathrm{M}^{\substack{\text{principal}\\ \text{strata}} \rightarrow \substack{\text{cohort}\\ \text{data}}}\right)^{-1}$ can be broken down into two steps. The first involves the estimation of s_a, s_p, and s_n by invoking the assumptions. That is, by ignorability:

$$s_a = P(S = a|G = 1) = P(S = a|G = 0) = e^{(1)}_{g=0}, \tag{A.5}$$

$$s_n = P(S = n|G = 0) = P(S = a|G = 1) = e^{(0)}_{g=1}, \quad \text{and} \tag{A.6}$$

$$s_p = 1 - s_a - s_n \tag{A.7}$$

Next, one may consider the quantities $P(Y = 1|S = a)$ and $P(Y = 1|S = n)$. By the exclusion restriction,

$$P(Y = 1|S = a) = P(Y = 1|S = a, G = 0) = y^{(1)}_{g=0,e=1}, \quad \text{and} \tag{A.8}$$

$$P(Y = 1|S = n) = P(Y = 1|S = n, G = 1) = y^{(1)}_{g=1,e=0} \tag{A.9}$$

Finally, the rearrangement of equations (A.1)-(A.2) yields $P(Y(g) = 1|S = p)$, for $g = 0, 1$.

Proof of the Result:

The result in Section 4 follows directly by applying the inverse mapping described above on the joint distribution $P^{\text{case}}_{\text{control}}(G, E, Y)$ in the case-control data. In order to write down these expressions, let us first consider the induced probabilities s_n, s_a, and s_p after case-control sampling. Then, from the previous section, we have:

$$
\begin{aligned}
s_n^{\overset{\text{case}}{\text{control}}} &\equiv P^{\text{case}}_{\text{control}}(S = n) \\
&= P^{\text{case}}_{\text{control}}(S = n | G = 1) \qquad \text{by (A.6)} \\
&= \frac{\sum_{y'} P(G = 1, E = 0 | Y = y') P^{\text{case}}_{\text{control}}(Y = y')}{\sum_{y', e'} P(G = 1, E = e' | Y = y') P^{\text{case}}_{\text{control}}(Y = y')} \\
&= \frac{\sum_{y'} P(G = 1, E = 0, Y = y') \cdot \frac{P^{\text{case}}_{\text{control}}(Y=y')}{P(Y=y')}}{\sum_{y', e'} P(G = 1, E = e', Y = y') \cdot \frac{P^{\text{case}}_{\text{control}}(Y=y')}{P(Y=y')}} \\
&= \frac{\sum_{y'} y^{(y')}_{g=1,e=0} \cdot e^{(0)}_{g=1} \cdot g \cdot \frac{P^{\text{case}}_{\text{control}}(Y=y')}{P(Y=y')}}{\sum_{y', e'} y^{(y')}_{g=1,e=e'} \cdot e^{(e')}_{g=1} \cdot g \cdot \frac{P^{\text{case}}_{\text{control}}(Y=y')}{P(Y=y')}}, \\
s_a^{\overset{\text{case}}{\text{control}}} &= P^{\text{case}}_{\text{control}}(S = a | G = 0) \\
&= \frac{\sum_{y'} P(G = 0, E = 1 | Y = y') P^{\text{case}}_{\text{control}}(Y = y')}{\sum_{y', e'} P(G = 0, E = e' | Y = y') P^{\text{case}}_{\text{control}}(Y = y')} \\
&= \frac{\sum_{y'} y^{(y')}_{g=0,e=1} \cdot e^{(1)}_{g=0} \cdot (1 - g) \cdot \frac{P^{\text{case}}_{\text{control}}(Y=y')}{P(Y=y')}}{\sum_{y', e'} y^{(y')}_{g=0,e=e'} \cdot e^{(e')}_{g=0} \cdot (1 - g) \cdot \frac{P^{\text{case}}_{\text{control}}(Y=y')}{P(Y=y')}}, \qquad \text{and} \\
s_p^{\overset{\text{case}}{\text{control}}} &= 1 - s_a^{\overset{\text{case}}{\text{control}}} - s_n^{\overset{\text{case}}{\text{control}}}
\end{aligned}
$$

where $g = P(G = 1)$. Now, from equations (A.1) and (A.2), we have that:

$$
\begin{aligned}
P^{\text{case}}_{\text{control}}(Y(0)|S=c) &= \left[\frac{P(G=0,E=0|Y=y)P^{\text{case}}_{\text{control}}(Y=y)}{\sum_{y',e'} P(G=0,E=e'|Y=y')P^{\text{case}}_{\text{control}}(Y=y')}\right. \\
&\qquad \left. - \frac{P(G=1,E=0|Y=y)P^{\text{case}}_{\text{control}}(Y=y)}{\sum_{y',e'} P(G=1,E=e'|Y=y')P^{\text{case}}_{\text{control}}(Y=y')}\right] \Big/ s_p^{\substack{\text{case}\\ \text{control}}} \\
&= \left[\frac{y^{(1)}_{g=0,e=0}\cdot e^{(0)}_{0}\cdot(1-g)\cdot\frac{P^{\text{case}}_{\text{control}}(Y=y)}{P(Y=y)}}{\sum_{y',e'} y^{(y')}_{g=0,e=e'}\cdot e^{(e')}_{0}\cdot(1-g)\cdot\frac{P^{\text{case}}_{\text{control}}(Y=y')}{P(Y=y')}}\right. \\
&\qquad \left. - \frac{y^{(1)}_{g=1,e=0}\cdot e^{(0)}_{g=1}\cdot g\cdot\frac{P^{\text{case}}_{\text{control}}(Y=y)}{P(Y=y)}}{\sum_{y',e'} y^{(y')}_{g=1,e=e'}\cdot e^{(e')}_{g=1}\cdot g\cdot\frac{P^{\text{case}}_{\text{control}}(Y=y')}{P(Y=y')}}\right] \Big/ s_p^{\substack{\text{case}\\ \text{control}}}
\end{aligned}
\tag{A.10}
$$

And,

$$
\begin{aligned}
P^{\text{case}}_{\text{control}}(Y(1)|S=c) &= \left[\frac{P(G=1,E=1|Y=y)P^{\text{case}}_{\text{control}}(Y=y)}{\sum_{y',e'} P(G=1,E=e'|Y=y')P^{\text{case}}_{\text{control}}(Y=y')}\right. \\
&\qquad \left. - \frac{P(G=0,E=1|Y=y)P^{\text{case}}_{\text{control}}(Y=y)}{\sum_{y',e'} Pr(G=0,E=e'|Y=y')P^{\text{case}}_{\text{control}}(Y=y')}\right] \Big/ s_p^{\substack{\text{case}\\ \text{control}}} \\
&= \left[\frac{y^{(1)}_{g=1,e=1}\cdot e^{(1)}_{g=1}\cdot g\cdot\frac{P^{\text{case}}_{\text{control}}(Y=y)}{P(Y=y)}}{\sum_{y',e'} y^{(y')}_{g=1,e=e'}\cdot e^{(e')}_{g=1}\cdot g\cdot\frac{P^{\text{case}}_{\text{control}}(Y=y')}{P(Y=y')}}\right. \\
&\qquad \left. - \frac{y^{(1)}_{g=0,e=1}\cdot e^{(1)}_{g=0}\cdot(1-g)\cdot\frac{P^{\text{case}}_{\text{control}}(Y=y)}{P(Y=y)}}{\sum_{y',e'} y^{(y')}_{g=0,e=e'}\cdot e^{(e')}_{g=0}\cdot(1-g)\cdot\frac{P^{\text{case}}_{\text{control}}(Y=y')}{P(Y=y')}}\right] \Big/ s_p^{\substack{\text{case}\\ \text{control}}}
\end{aligned}
$$

There is no further general simplification of the above. Hence, as these forms differ from the standard IV forms when $P^{\text{case}}_{\text{control}}(Y=y')$ deviates from $P(Y=y')$, the result follows. Alternatively, one may consider the example in Table (4.1) which shows this result by means of a counter-example.

References

[1] J. Angrist, J. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *JASA*, 91(434):444–455, 1996.

[2] C.S. Carlson, S.F. Aldred, P.K. Lee, R.P. Tracy, S.M. Schwartz, M. Rieder, K. Liu, O.D. Williams, C. Iribarren, E.C. Lewis, et al. Polymorphisms within the C-reactive protein (CRP) promoter region are associated with plasma CRP levels. *The American Journal of Human Genetics*, 77(1):64–77, 2005.

[3] G. Davey Smith and S. Ebrahim. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22, 2003.

[4] V. Didelez and N. Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309, 2007.

[5] T.P. Erlinger, E.A. Platz, N. Rifai, and K.J. Helzlsouer. C-reactive protein and the risk of incident colorectal cancer. *JAMA*, 291(5):585–590, 2004.

[6] C.E. Frangakis and D.B. Rubin. Principal stratification in causal inference. *Biometrics*, 58:21–29, 2002.

[7] National Cancer Institute. Surveillance epidemiology and end results.

[8] RL Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

[9] N.J. Timpson, D.A. Lawlor, R.M. Harbord, T.R. Gaunt, I.N.M. Day, L.J. Palmer, A.T. Hattersley, S. Ebrahim, G.D.O. Lowe, A. Rumley, et al. C-reactive protein and its role in

metabolic syndrome: mendelian randomisation study. *The Lancet*, 366(9501):1954–1959, 2005.

[10] K.K. Tsilidis, K.J. Helzlsouer, M. W. Smith, V. Grinberg, Hoffman-Bolton J., Clipp S. L., Visvanathan K., and E.A. Platz. Association of common polymorphisms in il10, and in other genes related to inflammatory response and obesity with colorectal cancer. *Cancer Causes Control*, 2009 [accepted].