The International Journal of Biostatistics

Volume 8, *Issue* 1 2012 *Article* 2

Designs Combining Instrumental Variables with Case-Control: Estimating Principal Strata Causal Effects

Russell T. Shinohara, *Johns Hopkins University*
Constantine E. Frangakis, *Johns Hopkins University*
Elizabeth Platz, *Johns Hopkins University*
Konstantinos Tsilidis, *University of Ioannina*

Recommended Citation:
Shinohara, Russell T.; Frangakis, Constantine E.; Platz, Elizabeth; and Tsilidis, Konstantinos (2012) "Designs Combining Instrumental Variables with Case-Control: Estimating Principal Strata Causal Effects," *The International Journal of Biostatistics*: Vol. 8: Iss. 1, Article 2.
DOI: 10.2202/1557-4679.1355

Designs Combining Instrumental Variables with Case-Control: Estimating Principal Strata Causal Effects

Russell T. Shinohara, Constantine E. Frangakis, Elizabeth Platz, and Konstantinos Tsilidis

Abstract

The instrumental variables framework is commonly used for the estimation of causal effects from cohort samples. However, the combination of instrumental variables with more efficient designs such as case-control sampling requires new methodological consideration. For example, as the use 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 variables analysis does not appropriately estimate the causal effects of interest when the instrumental variables design is combined with the case-control design. We also propose a method that can estimate the causal effects in such combined designs. We illustrate the method with a study in oncology.

KEYWORDS: case-control, instrumental variables, Mendelian randomization, principal stratification, study design

Author Notes: The authors thank Nicholas Jewell and two reviewers for constructive comments, Jeff Leek and Hongkai Ji for discussions on properties of genetic recombinations, and the NIH (R01 DA023879) for partial support.

1 Introduction

An increasingly common methodology in epidemiology that uses a genetic variable to assess the effect of an exposure on an outcome is known as Mendelian randomization. Indeed, these techniques have been studied by a number of authors and several extensive reviews and discussions are available in the medical and epidemiological literature (Didelez and Sheehan, 2007, Davey Smith and Ebrahim, 2003, Davey Smith and Ebrahim, 2004, Davey Smith et al., 2005, Davey Smith, 2006, Davey Smith et al., 2008, and Lawlor et al., 2008). Whereas the relationship between the exposure E and outcome Y may be confounded, the mechanism by which the genotype G is assigned is well-understood and stems from meiosis and fertilization. This random process allows for the estimation of the causal effect of E on Y through instrumental variable techniques (Angrist et al., 1996).

An example of such a 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 here is the effect that chronically elevated C-reactive protein levels (CRP) have on the risk of colorectal cancer. CRP concentration was measured in archived baseline blood samples. For cost reasons common to many such studies, the genetic variables were measured only for a nested case-control sample of 172 colorectal cancer cases and 342 matched controls were selected based on age, sex, and date of blood draw.

For instrumental variables estimation using designs alternative to a cohort, early work by Mauro (2007) in economics used a design that matched firms that were *exposed vs. unexposed* to the treatment of interest. The work of Didelez and Sheehan (2007) investigated Mendelian randomization in case-control studies. They noted that external information is required in order to estimate bounds for causal effects. They found that for testing the null hypothesis of no causal effect the composite design does not pose a problem. Our interest is in local treatment effects (Angrist et al.), and we consider the estimation of such effects from data arising from a composite design.

We show that the methodology of standard instrumental variables is not directly applicable to estimate general causal effects when instrumental variables are combined with case-control designs. This is because (a) case-control data alone do not include any information about the prevalence or incidence of disease in the population (our points apply when either prevalence or incidence over a time period, as in the colorectal cancer study, is involved); and (b) for estimation of effects with instrumental variables using case-control data such incidence information is important, in contrast to estimation of simple odds ratios of association, where such incidence information is ancillary (Prentice and Pyke, 1979).

We also show how to address this problem. First, as the incidence information is not available from the case-control study alone, it must be obtained externally. Second, this information must be combined with the observed case-control data in order to reconfigure an estimate of the cohort distribution of the full data. In that reconfiguration, the instrumental variable analysis can be applied to yield appropriate estimates of the effects.

Among instrumental variables applications, Mendelian randomization is a special case. Mendelian randomization has limitations. We discuss the plausibility of the assumptions in the methodology section. We also discuss the utility of our methods for more general instrumental variables. Importantly, the design idea in this paper is not just about Mendelian randomization, but about any physical setting for which instrumental variables can be plausible. This includes many fields such as, for example, using access to special hospitals as an instrument to study myocardial infarction outcomes (Newhouse and McClellan, 1998), or using calendar variation of hormone replacement therapy rates as instrument to study the effect of its use on cardiovascular outcomes (Shetty et al, 2009); see also Hernan and Robins (2006) for a review. We hope that our results on case-control designs with instrumental variables will stimulate and facilitate the use of instrumental variables when more flexible designs other than simple cohorts are needed.

In the next section, we review the instrumental variables frameworks and give notation to define the estimation target. We then describe the composite case-control instrumental variables design in detail and suggest an approach for estimation. In Section 2.4, we consider the assumptions necessary for the standard analysis of Angrist et al. (1996) and thus our analysis, and then we briefly review the cohort analysis methodology. We then give the main result stating that the standard cohort instrumental variables analysis is not appropriate for the analysis of composite designs. We note that our method provides a correct estimation procedure and conclude with a discussion.

2 Methods

2.1 Frameworks of instrumental variables

Several frameworks are available through which estimation can be conducted using instrumental variables. Vytlacil (2002) demonstrated that the structural equation paradigm is equivalent to that of potential outcomes and the equivalence of graphical causal models to these setups was shown by Pearl (2000). We proceed within the potential outcome framework to show our points, noting that these equivalences indicate that our results apply more generally to the other frameworks.

2.2 Notation and estimation target

We wish to estimate the effect of chronically elevated CRP on colorectal cancer using the SNP rs1205 as an instrumental variable (details on the choice of this SNP and its relationship to the quantities of interest are published elsewhere, Tsilidis et al., 2009)). The answer can inform about the existing question of whether C-reactive protein has a causal role in cancer risk or if it is merely a surrogate of other processes involved in inflammation that can have a causal role in cancer risk (Allin et al., 2009). To do this, first consider a population of units i representable by the original *cohort* sample (for example, the CLUE II cohort sample). For each individual the genotype g takes its value at meiosis and fertilization (time 0 of Figure 1) and can be either $g = 0$ (the rs1205 genotype (CT/TT) associated with less CRP) or $g = 1$ (the genotype (CC) associated with more CRP). By this notation, we intend that the observed locus g be either the locus of a causal agent for E or close enough to such a causal agent that recombinations between the two loci are unlikely (see Section 3, paragraph 3). Let $E_i(g)$ be the level of CRP (1 for higher than the median; 0 for lower than the median; median=2.12 mg/L) that the i-th individual would experience at a later time 1 if the genotype were g. Further, let $Y_i(g)$ denote the colorectal cancer status of the i-th individual at a later time 2 if the genotype were g. These variables here are binary, but the main points of the paper are generalizable to more complex types; see our Section 3.

In this population cohort we make the usual assumptions for using the genotype as an instrumental variable, in the sense of Angrist et al., and which are scientifically plausible in our case (see section 2.2 for a discussion of these assumptions).

We wish to use the genetic variable to assess the effect that changing CRP (E) has on colorectal cancer. If variations of the SNP (G) have no effect on cancer other than when changing CRP (see also the exclusion restriction later), then the effect that changing E has on cancer occurs only for those individuals for whom G changes E. For this reason we set the target quantity of interest to be the effect that G has on cancer for the individuals for whom G does increase E:

$$P(Y_i(1) \mid E_i(1) - E_i(0) = 1) \text{ versus } P(Y_i(0) \mid E_i(1) - E_i(0) = 1) \qquad (1)$$

To see that formula (1) describes the effect spelled out above, note that the "subjects for whom the genotype does increase CRP expression" are those for whom CRP expression under genotype 0 would be low, namely $E_i(0) = 0$, but CRP expression under genotype 1 would be high, namely $E_i(1) = 1$; combining these we get that these are the subjects for whom $E_i(1) - E_i(0) = 1$. Note that this group of patients is by nature not identifiable from the observed data, but nonetheless the target (1) is estimable under assumptions described in Angrist et al. In the remainder of the

paper, we focus on estimation of $P(Y_i(g) \mid E_i(1) - E_i(0) = 1)$ for $g = 0, 1$, from which the causal risk difference, odds ratio, and relative risk may be estimated.

2.3 Study design and estimation

The composite case-control instrumental variables design is summarized in Figure 1. For each participant, let G_i denote the *actual* value that the genotype takes at meiosis and fertilization (time 0). Let $E_i(= E_i(G_i))$ denote the actual value that the circulating C-reactive protein concentration takes at time 1. At this time, blood is drawn from each participant and stored for possibly measuring these values, E_i and G_i, depending on later information available at time 2.

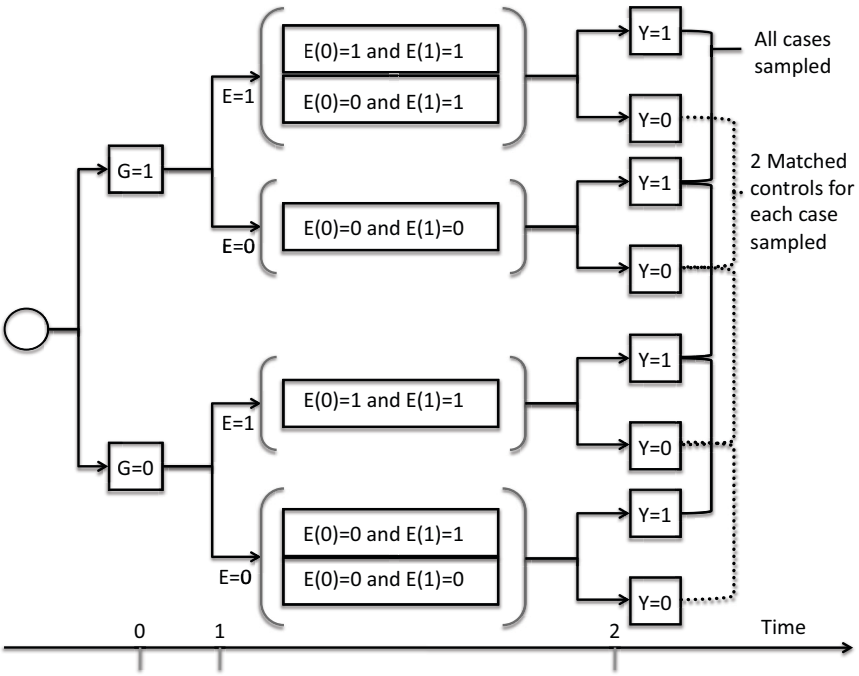

Figure 1: Summary of the design of case-control sampling in the context of instrumental variables.

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. The observed data therefore are $\{X_i, G_i, E_i\}$ for all cases and an $X-$ matched sample of controls.

To estimate the target (1) correctly from our composite study design, we make two observations: The first is that under instrumental variables assumptions (discussed in the next section) in the cohort, the standard instrumental variables methodology (Angrist et al., and given in Appendix A) gives correct answers if applied to the *cohort joint* distribution of the genotypes, expression and outcome, $P(G_i, E_i, Y_i \mid X_i)$, conditionally on the covariates. It does not, however, generally give correct answers if applied to the composite design as we show in Section 2.4. The second observation is that the case-control design provides the part of the cohort distribution that is conditional on the outcome and matching covariates, namely $P(G_i, E_i \mid Y_i, X_i)$. Thus, we propose to address the problem in the following two stages:

1. Estimate a re-configuration of the *cohort joint* distribution of the genotypes, expression and cancer outcome, $P(G_i, E_i, Y_i \mid X_i)$, conditionally on the covariates. To re-configure $P(G_i, E_i, Y_i \mid X_i)$, note that

$$P(G_i, E_i, Y_i \mid X_i) = P(G_i, E_i \mid Y_i, X_i) \cdot P(Y_i \mid X_i)$$

   Thus, by the second observation above, to estimate the distribution $P(G_i, E_i, Y_i \mid X_i)$ we may multiply an estimate of $P(G_i, E_i \mid Y_i, X_i)$ from the observed case-control data with an external estimate of the disease incidence in the population $P(Y_i \mid X_i)$.
2. Apply the instrumental variables methodology (Angrist et al.) to the re-configured distribution $P(G_i, E_i, Y_i \mid X_i)$. The resulting formulae are given in Appendix B.

The concept behind the above estimation scheme is more general. When covariates can be measured in a cohort with instrumental variables, a number of alternative designs is generally possible. The two-stage estimation used in this paper suggests that there is a generalizable approach to use such designs: first, place instrumental variables assumptions on a cohort when plausible; second, find what extra information (e.g., the marginal incidence with the case-control design) is required to supplement to the design data in order to re-configure an estimate of the cohort distribution of the data. Finally, apply the instrumental variables methodology to the re-configured cohort distribution to estimate the causal effects of interest. To our knowledge this more general fact has not been realized for Mendelian randomization or other instrumental variables problems.

In the next section, we examine the instrumental variable assumptions from Angrist et al. carefully in the context of our Mendelian randomization study. We then proceed with a brief review of the standard cohort instrumental variables before giving the main result of this work, which states in more detail the properties of the two-step estimator from this section.

2.4 Instrumental variables assumptions

We consider the study cohort at time 0 to be a simple random sample of the population cohort of interest (as in the above section defining the scientific target), so that $(X_i, G_i, E_i(g), Y_i(g))$ are independent and identically distributed from that population where X_i are pertinent covariates, including those used for matching. 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., and that are scientifically plausible in our case. For brevity, we use the label of S_i for the vector $(E_i(0), E_i(1))$ and call S_i the principal strata.

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

ASSUMPTION 1. *Ignorability*

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

For Mendelian randomization, ignorability is supported by the random receipt of a paternal vs. maternal allele at fertilization. Assumption 1 is typically violated if one of the following two subconditions is violated: first, if there is no population admixture; second, if the variation in G that is caused by the meiotic process is independent on the variation in other causal agents that act on Y through pathways other than changing E. To simplify notation and with no loss of generality, we suppose that we are within a particular level of the covariates unless otherwise noted.

A second assumption is the absence of any effect of the rs1205 genotype on colorectal cancer if the genotype has no effect on CRP; 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 mechanisms (Timpson et al., 2005). An important note is that in many scenarios,

the coarsening or dichotomization of the exposure E_i may be problematic. In our example, the assumption is supported by our scientific knowledge that CRP is a measure which is noisy and relates to the patient's health status on the particular date of the blood draw (for example, CRP may be affected by a minor bruise). Severe violations of any of the assumptions can sometimes be detected by negative estimates of probabilities, but this did not happen in the CLUE study.

The final assumption is that there are no individuals in the study who would suffer from high CRP under the genotype associated with lower CRP but from low CRP under the alternate genotype. That is, there are no discordant individuals:

ASSUMPTION 3. *Monotonicity*

$$E_i(g=0) \leq E_i(g=1)$$

Monotonicity should be used only when the SNPs have a well understood effect on the exposure of interest. This will be the case when the variant of the specific SNP disrupts the usual transcription of the gene, and that disruption is for a reason specific to that variant. For rs1205, violations of monotonicity here would be inconsistent with the established biological mechanism by which the SNP g predisposes to CRP (as above for exclusion restriction). Under monotonicity, we relabel $S_i = a$ for individuals with $(E_i(0) = 1$ and $E_i(1) = 1))$; $S_i = p$ for individuals with $(E_i(0) = 0$ and $E_i(1) = 1)$; and $S_i = n$ for individuals with $(E_i(0) = 0$ and $E_i(1) = 0)$.

2.5 Review of instrumental variables in cohort studies

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.

Under the above assumptions of instrumental variables, Angrist et al. show that the causal effect (1) is estimable using data arising from a cohort design. To be more specific, suppose here we are already within cells of the covariates X. Then Angrist et al.'s main result is that there is a bijection between the cohort data $P(G,E,Y)$ and the potential outcomes $(P(G),P(S),P(Y(g) \mid S))$ of interest. More details concerning this work are provided in Appendix A and a proof is included for convenience.

If the true contrast (1) is null, this cohort instrumental variables methodology also gives no effect when applied naively to case-control data. In the next section, however, we show that when a non-zero effect is present the design-naive instrumental variables methodology does not yield correct results if applied to data

arising from a composite case-control instrumental variables design. This is in contrast to the standard odds ratio estimation for association, which is known to be invariant to case-control sampling.

2.6 Methodological implications of design on analysis

We now consider what happens when we have collected data from the above composite design of case-control and instrumental variables. In Section 2.3, we suggest a two-step procedure for estimating causal effects. The value of this procedure is described in our main result:

Result 1. Suppose the instrumental variables assumptions 1-3 hold in the cohort.

1. If the cohort instrumental variables methodology is applied to even an increasingly large sample of data from the composite study design described above to estimate the causal effect (1), the results will generally differ from the true value of (1). This discrepancy holds even if the effect is an odds ratio.
2. The true value of (1) is identifiable if one also has the marginal distribution of Y given covariates in the cohort.

We prove Result 1.2 in Appendix B by deriving the target quantities (1) adjusting for the marginal distribution of the outcome in the cohort. Result 1.1 is shown through a counterexample below. The forms given in Appendix B are also especially useful for applications, as they may be used directly for the analysis of data arising from composite designs.

One way of understanding this result is the following. The case-control sampling leaves invariant the usual odds ratio between the outcome variable and a factor of interest, that is, one type of association between two observed variables. However, the instrumental variables effect (1) is an effect describable within a subset of individuals not directly observed. Specifically, to describe this effect in terms of observed variables, one must decompose a mixture into components that are not directly observed but that are derivable via Assumptions 1-3 (Lemma 1). Thus, since this effect is not describable as an association between the outcome and an observed variable, it is understandable that it does not share the invariance property under case control sampling.

From the formulas derived in Appendix B and after some additional algebra (omitted), it can be shown that the magnitude of the bias depends on two intuitive factors. First, if the true target effect (1) is null, then the bias will be null, and the bias generally increases with the magnitude of the true effect. The second factor that influences bias in the odds ratio effect is the magnitude of confounding that is addressed by the instrumental variable in comparison to a standard regression; the

smaller the confounding, the smaller the bias. The reason for this is that under the usual condition of no confounding, each of the causal odds ratio estimands in the cohort and the composite designs are equal to their association counterparts, which are the same by the invariance property of odds ratios in case-control studies.

The confounding mentioned above is the comparison of the outcome distribution across different principal strata S in the cohort, for the same observed value of instrument and exposure, and is assessed by an indirect argument. For example, consider the comparison of $P(Y \mid E = 0, G = 1)$ to $P(Y \mid E = 0, G = 0)$ in the cohort. The former is actually also equal to $P(Y \mid S = n, E = 0, G = 0)$, and the latter is a mixture between $P(Y \mid S = n, E = 0, G = 0)$ and $P(Y \mid S = p, E = 0, G = 0)$ (this follows from Appendix A). Thus an observed difference between $P(Y \mid E = 0, G = 1)$ to $P(Y \mid E = 0, G = 0)$ implies that the non-directly observed strata $S = n$ and $S = p$ have different outcomes when having the same instrument and exposure (0).

To demonstrate the second factor more clearly, we provide an illustration: Table 1 shows both the correct effects and the answers produced by the cohort instrumental variables methodology applied directly, which we call "design-naive", in a variety of settings.

The settings are chosen by setting values for the components of $P(S)$ and $P(Y(g) \mid S)$, and then calculating the true effect (1) and the design-naive values from ignoring the composite design. The values for $P(S)$ are chosen to be $\frac{1}{3}$ each for simplicity. The values of $P(Y(g) \mid S)$ are chosen so that there is an effect for $S = p$, that there is no confounding of the relationship between G and Y by $S = p$ and $S = a$ ($P(Y = 1 \mid S = a)$ is set to equal $P(Y(1) = 1 \mid S = p)$), but that there can be such confounding by $S = p$ and $S = n$ depending on whether we choose $P(Y = 1 \mid S = n)$ to be different from or equal to $P(Y(0) = 1 \mid S = p)$ (which is equal to $P(Y = 1 \mid S = p, E = 0, G = 0)$). Table 1 confirms that increased confounding increases the bias even for the odds ratio. In more extreme cases (not shown) when the confounding is severe, ignoring the design even produces some probability estimates outside the boundaries of [0,1].

2.7 Application: CLUE II cohort

We applied this estimation procedure to the CLUE II cohort that has been introduced in the above sections. The data are summarized in Table 2. In order to model the joint distribution of observables in the case-control study, first we fitted the conditional models corresponding to $P(E_i \mid Y_i, X_i)$ and $P(G_i \mid E_i, Y_i, X_i)$ via logistic regression. We then multiplied these fitted distributions, as in Part (i) above, to obtain the fitted distribution $P(G_i, E_i \mid Y_i, X_i)$. This distribution we multiplied by the SEER (2009) incidence estimates for $P(Y \mid X)$, to obtain the fitted distribution for $P(G_i, E_i, Y_i, X_i)$, also as in Part (i). Finally, for the quartile values of age

for men and women (covariates X), we applied the cohort instrumental variables methodology (from Angrist et al., given in Appendix A) onto the fitted distribution $P(G_i, E_i, Y_i, X_i)$ and thus obtained estimates of the distribution $P(E_i(0), E_i(1) \mid X_i)$ and $P(Y_i(g) \mid E_i(0), E_i(1), X_i)$, and hence also for the effects (1). The standard errors were estimated via the delta method on the resulting composite mapping from the logistic models for $P(E_i \mid Y_i, X_i)$, $P(G_i \mid E_i, Y_i, X_i)$ and the SEER estimates $P(Y \mid X)$ to the distributions $P(E_i(0), E_i(1) \mid X_i)$ and $P(Y_i(g) \mid E_i(0), E_i(1), X_i)$.

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

(i) Distribution of Target Quantities in Population:

	g=0		g=1
$P(S=a)^{(a)}$ = 1/3	$P(Y=1\|S=a)$ = 4‰	(=)	$P(Y=1\|S=a)$ = 4‰
$P(S=p)^{(a)}$ = 1/3	$P(Y(g=0)=1\|S=p)$ = 1‰		$P(Y(g=1)=1\|S=p)$ = 4‰
$P(S=n)^{(a)}$ = 1/3	$P(Y=1\|S=n)$	(=)	$P(Y=1\|S=n)^{\dagger}$
	$P(G=0)$ = 50%		$P(G=1)$ = 50%

(ii) True and design-naive values of causal effects after case-control design:
(comparing $P(Y(g=0)=1|S=p)$ to $P(Y(g=1)=1|S=p)$)

†when $P(Y=1\|S=n)$ is	1‰	2‰	4‰
then the true Difference is	0.3‰	0.3‰	0.3‰
and the naive Difference is	33% $(\mathbf{1.10 \times 10^3})^{(b)}$	29% $(\mathbf{9.95 \times 10^2})$	25% $(\mathbf{8.49 \times 10^2})$
the true Odds Ratio is	4	4	4
and the naive Odds Ratio is	4 **(1.00)**	3.5 **(0.88)**	2.96 **(0.74)**
the true Relative Risk is	4	4	4
and the naive Relative Risk is	2.15 **(0.54)**	2.04 **(0.51)**	1.91 **(0.48)**

$^{(a)}$ We use the label of $S_i = a$ for individuals with $(E_i(0) = 1$ and $E_i(1) = 1)$; the label of $S_i = p$ individuals with $(E_i(0) = 0$ and $E_i(1) = 1)$; and the label of $S_i = n$ for individuals with $(E_i(0) = 0$ and $E_i(1) = 0)$.

$^{(b)}$ The ratios of design-naive to true values are shown in bold face font in parentheses.

The estimation of the target quantities in the colorectal oncology study was conducted with and without addressing the case-control design. As these data were part of a pilot study and the sample size is relatively small, the estimated standard errors were large (95% confidence intervals included the null values). Nevertheless, the point estimates demonstrate the possible differences that are plausible between the results of the two methods in a real problem.

In this case, the odds ratios between the two methods are almost the same, which is consistent with a scenario that there may be no considerable confounding in the CRP 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.

3 Discussion

We addressed the combination of case-control and instrumental variables designs. This is of interest in itself, but is especially important as the number of Mendelian randomization studies grows. As genetic data is often 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 a 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.

Mendelian randomization has its limitations. For one, its assumptions may be questionable, and their plausibility, as discussed in the paper, should be contemplated. Also, the utility of Mendelian randomization more generally is directly related to the strength by which a SNP is associated to the intermediate exposure and, ultimately, to the clinical outcome. Here, the application to the CLUE study gives a sense of the numerical discrepancies that can arise. Appropriate extensions can then be used in larger and more challenging studies. An example is gout which has been shown to be strongly related to a set of SNPs (Dehghan et al., 2008). These extensions will be examined in future work.

It is important to note that when we refer to the effect of rs1205 on exposure and outcome, we mean the effect of the meiotic process that leads to the variation of rs1205. For the treatment of this paper, as also described by Joffe (2011), we do not actually require that the SNP is the ultimate causal agent itself although we do require that the SNP and causal agent locus are close enough that the chance of recombination is negligible. In our problem, rs1205 is in the CRP gene and the recombination rate in this region in approximately 2.2 cM/Mb (Kong et al., 2002). In the 2kb region that encompasses the CRP gene, this corresponds to an estimated recombination probability of less than 0.0001. If recombination is more likely, however, then the observed SNP can be considered as a surrogate of the desired instrument - the causal gene. Hernan and Robins (2006, Theorem 5) show that if, in such a case, monotonicity still holds for the causal gene, then the difference causal effect (i.e., the difference in averages of the distributions in (1) with the observed

SNP replaced with the causal gene), can still be estimated by using the observed SNP as if it were the causal gene. For non-linear estimands though, such as relative risks or odds ratios, the problem appears to be more complex and needs additional treatment.

Table 2. Description of the CLUE II data.

	Mean in study (n=509)	**Mean in Cases** (171 subjects)	**Mean in Controls** (338 subjects)	**Difference** (SE)
Age (SD)	63.45 (11.3)	63.55 (11.3)	63.40 (11.3)	0.15 (1.06)
Sex (% Male)	0.44	0.44	0.44	0.00 (0.05)
RS1205 Genotype (% with genotype CC)	0.46	0.47	0.45	0.02 (0.05)
CRP (% with CRP larger than 2.115 mg/L)	0.50	0.55	0.48	0.07 (0.05)

Median CRP: [25^{th} %, 75^{th} %] (in mg/L)	**in study** (n=509)	**in (CC/TT) Genotype Group**	**in (CC) Genotype Group**	**in Low CRP Group** ($\leq$ 2.12 mg/L)	**in High CRP Group** ($>$ 2.12 mg/L)
Men at:					
Age 55	1.74 [0.88, 3.45]	1.58 [0.80, 3.11]	2.00 [1.02, 3.95]	0.99 [0.59, 1.47]	3.91 [2.81, 6.16]
Age 66	2.05 [1.04, 4.06]	1.86 [0.95, 3.67]	2.37 [1.20, 4.66]	1.06 [0.65, 1.52]	4.14 [2.90, 6.66]
Age 72	2.24 [1.13, 4.44]	2.04 [1.04, 4.02]	2.59 [1.32, 5.11]	1.10 [0.68, 1.55]	4.28 [2.69, 6.98]
Women at:					
Age 55	1.92 [0.98, 3.82]	1.73 [0.88, 3.41]	2.20 [1.12, 4.33]	1.03 [0.63, 1.50]	4.05 [2.78, 6.46]
Age 66	1.96 [0.99, 3.89]	1.75 [0.89, 3.43]	2.22 [1.13, 4.36]	1.04 [0.64, 1.51]	4.07 [2.87, 6.52]
Age 72	1.98 [1.00, 3.93]	1.75 [0.89, 3.45]	2.23 [1.13, 4.38]	1.04 [0.64, 1.51]	4.09 [2.88, 6.56]

Abbreviations: SD, standard deviation; SE, standard error

Table 3. Design-based and design-naive estimates of the causal effects for median and quartile ages and each sex. 95% confidence intervals are based on the bootstrap.

	Men		Women	
	Design-based	Design-naive	Design-based	Design-naive
Age 55				
P(Y(0)=1|S=p)	6.47×10^{-4} 0, 4.46×10^{-3}	0.270 0, 1	4.43×10^{-4} 0, 3.38×10^{-3}	0.270 0, 1
P(Y(1)=1|S=p)	11.90×10^{-4} 0, 34.10×10^{-4}	0.403 0, 0.968	8.08×10^{-4} 0, 2.32×10^{-3}	0.402 0, 1
Difference	5.46×10^{-4} -3.41×10^{-3}, 4.46×10^{-3}	0.133 -0.968, 1	3.66×10^{-4} -2.32×10^{-3}, 3.38×10^{-3}	0.132 -1, 1
Odds Ratio	1.85 0, ∞	1.83 0, ∞	1.83 0, ∞	1.82 0, ∞
Relative Risk	1.84 0, ∞	1.49 0, ∞	1.83 0, ∞	1.49 0, ∞
Age 66				
P(Y(0)=1|S=p)	1.56×10^{-3} 0, 1.38×10^{-2}	0.270 0, 1	1.05×10^{-3} 0, 6.62×10^{-3}	0.269 0, 1
P(Y(1)=1|S=p)	2.84×10^{-3} 0, 5.50×10^{-3}	0.402 0, 0.731	1.92×10^{-3} 0, 4.73×10^{-3}	0.401 0, 0.814
Difference	1.28×10^{-3} -5.50×10^{-3}, 1.38×10^{-2}	0.132 -0.731, 1	8.70×10^{-4} -4.73×10^{-3}, 6.62×10^{-3}	0.132 -0.814, 1
Odds Ratio	1.82 0, ∞	1.82 0, ∞	1.83 0, ∞	1.82 0, ∞
Relative Risk	1.82 0, ∞	1.49 0, ∞	1.83 0, ∞	1.49 0, ∞
Age 72				
P(Y(0)=1|S=p)	2.12×10^{-3} 0, 1.99×10^{-2}	0.269 0, 1	1.41×10^{-3} 0, 9.15×10^{-3}	0.269 0, 1
P(Y(1)=1|S=p)	3.85×10^{-3} 0, 8.64×10^{-3}	0.403 0, 0.823	2.58×10^{-3} 0, 7.47×10^{-3}	0.400 0, 1
Difference	1.73×10^{-3} -8.64×10^{-3}, 1.99×10^{-2}	0.134 -0.823, 1	1.17×10^{-3} -7.47×10^{-3}, 9.15×10^{-3}	0.132 -1, 1
Odds Ratio	1.82 0, ∞	1.83 0, ∞	1.83 0, ∞	1.82 0, ∞
Relative Risk	1.82 0, ∞	1.5 0, ∞	1.83 0, ∞	1.49 0, ∞

Our work is related to suggestions that certain causal effects of interest are not identifiable from such combined designs, but that they can be identified by supplementing the missing information about the incidence of the outcome (Li and Frangakis, 2006, and Constantinou, 2009). These works, however, do not directly address our problem. Specifically, Constantinou conceptualizes causal effects through outcomes that would have been observed if every person had been forced to have high (and low) CRP. We do not consider such outcomes here because they are not all potentially observable (Rubin, 1974) in the study: neither the instrument

nor any other mechanism is known to fully control CRP. Moreover, Li and Frangakis's discussion merely suggested the theoretical possibility of using instrumental variables with other designs, but did not specifically address the problem.

Some estimands of instrumental variables using other approaches require that one conceptualizes certain other outcomes. For example, for a subject who would have high expression under both genotypes (i.e., $E_i(1) = E_i(0) = 1$), the approach of Bowden and Vansteelandt Bowden and Vansteelandt (2011) [1] relies on contemplating the cancer outcome that this subject would have if in some way they had had *low* CRP expression. The problem with these approaches is that in this study, there is no such mechanism to obtain information about such outcomes, and so they are ill defined: there could perhaps be even many ways to have low CRP but none of which is acting as such for this subject in this study. It is for this reason that we focus on estimating the goal in (1).

After a method is used to estimate the effects, the uncertainty can be estimated by various approaches including the bootstrap or delta method. The bootstrap for a method that adjusts for the matching covariates can sample with replacement directly individuals from the case-control population, within levels of Y, to reflect the uncertainty of estimating the parameters in the distributions $P(G,E \mid Y,X)$ (this follows from the ignorability results of Rubin, 1978). In our example, the estimate of the population prevalence $P(Y \mid X)$ is assumed known without sampling error, but in other cases one should include uncertainty for that estimation as well.

If in relation to the sample size there are only few covariate levels to stratify for matching and to make Assumptions 1-3 more plausible, then the estimation can follow directly using saturated models. In most applications, though, as in our example, analysis will need to adjust more parsimoniously for multilevel or continuous matching covariates and for covariates that make Assumptions 1-3 more plausible. For these cases, a saturated approach may not be possible, and modeling of $P(E \mid Y,X)$ and $P(G \mid E,Y,X)$ may be needed. One limitation of this is that restrictions of such modeling in the observed data distributions induce restrictions on the effects (1) through the instrumental variables method described in Appendix A. This issue is also related to the null paradox (see, for example, Robins et al., 1999). This is the fact that a class of mispecified models for the components $P(E \mid Y,X)$ and $P(G \mid E,Y,X)$ may not contain a distribution that satisfies the null hypothesis of no causal effects. The extent to which this is a concern is dependent on the extent to which the models are incorrect. Such a problem due to marked mispecification can therefore be alleviated by model checking and model enriching for $P(E \mid Y,X)$ and $P(G \mid E,Y,X)$.

[1] The earlier, technical report version of this present paper, made public in 2009 in the report series of the authors' institution and available on request, precedes chronologically the paper by Bowden and Vansteelandt who cite that technical report.

DOI: 10.2202/1557-4679.1355

Our presentation focused on having dichotomized the instrument, exposure, and outcome. The binary instrument is presented here because circulating CRP level was similar between the groups with rs1205 genotypes (CT) and (TT), and to simplify the analysis. For multilevel instruments, extensions can be based, for example, on work developed in Frangakis et al. (2004) or Heckman et al. (2006), where, in the latter paper, if the original instrument has no natural ordering in terms of the likelihood of the exposure, a new instrumental variable is constructed that has such ordering. Our design corrections may also be applicable to methods with continuous exposure discussed by Vansteelandt and Goetghebeur (2003).

Under the assumptions we used, an alternative to the approach we took could be to model the target probabilities directly and estimate them using inverse probability weighting as has been used in other problems and models (Cole and Hernan, 2008, Abadie, 2003, and Tan, 2006). For example, we could solve the score equations that would have arisen if we had observed cohort data, the likelihood of which is described in Imbens and Rubin (1997), but weighted by the inverse of the probability of being selected into the case-control design. As solving the instrumental variables score equations directly is unstable even with cohort data, a modified EM algorithm would be needed to accomplish this, and will be explored in future work.

In conclusion, cost or efficiency considerations can suggest combining the case-control design with the instrumental variables design. With such composite designs, standard instrumental variables methods are not generally appropriate, and we have provided methods to better estimate the causal effects.

4 Appendix A: Instrumental variables estimation from a cohort design

The result of Angrist et al. may be summarized as:

Lemma 1. Under Assumptions 1-3

1. any given distribution of principal strata and potential outcomes in the cohort, along with a distribution of genotypes, induces a distribution on the cohort data $P(G,E,Y)$.
2. the mapping described in 1 is invertible; that is, we can use the distribution $P(G,E,Y)$ of the cohort data to find one and only one distribution for the principal strata and potential outcomes $(P(G),P(S),P(Y(g) \mid S))$ that gave rise to $P(G,E,Y)$.

Proof of Lemma 1 (as in Angrist et al.). We show the lemma by giving explicit forms for $P(G,E,Y)$ from $(P(G),P(S),P(Y(g) \mid S))$ and vice versa. Let us start with the function that maps a distribution $(P(G),P(S),P(Y(g) \mid S))$ to the distribution of the cohort data $P(G,E,Y)$. For the first distribution, abbreviate $P(S=s)$ by $S^{(s)}$, and $P(Y(g)=1 \mid S=s)$ by $Y^{(1)}_{G=g,S=s}$. For the second distribution, abbreviate $P(E=e|G=g)$ by $E^{(e)}_{G=g}$ and $P(Y=1|G=g,E=e)$ by $Y^{(1)}_{G=g,E=e}$. Then, one may note with the help of Figure 1 and the assumptions that

$$E^{(0)}_{G=0} = S^{(n)} + S^{(p)}, \quad E^{(1)}_{G=1} = S^{(a)} + S^{(p)}, \quad E^{(0)}_{G=1} = S^{(n)}, \quad E^{(1)}_{G=0} = S^{(a)}, \qquad \text{(A.1)}$$
$$\text{and so } S^{(p)} = 1 - E^{(0)}_{G=1} - E^{(1)}_{G=0}.$$

Also with the help of the mixtures in Figure 1 and Assumptions 1-3, we have that

$$Y^{(1)}_{G=1,E=0} = Y^{(1)}_{G=1,S=n}, \quad Y^{(1)}_{G=1,E=1} = \frac{S^{(p)}}{S^{(a)}+S^{(p)}} Y^{(1)}_{G=1,S=p} + \frac{S^{(a)}}{S^{(a)}+S^{(p)}} Y^{(1)}_{G=1,S=a}, \quad \text{and} \qquad \text{(A.2)}$$

$$Y^{(1)}_{G=0,E=1} = Y^{(1)}_{G=0,S=a}, \quad Y^{(1)}_{G=0,E=0} = \frac{S^{(p)}}{S^{(n)}+S^{(p)}} Y^{(1)}_{G=0,S=p} + \frac{S^{(n)}}{S^{(n)}+S^{(p)}} Y^{(1)}_{G=0,S=n} \qquad \text{(A.3)}$$

By noting that, from the exclusion restriction, $Y^{(1)}_{G=1,S=n} = Y^{(1)}_{G=0,S=n}$, equal, say, $Y^{(1)}_{S=n}$ and $Y^{(1)}_{G=1,S=a} = Y^{(1)}_{G=0,S=a}$, equal, say, $Y^{(1)}_{S=a}$, the recovery of $Y^{(1)}_{G=g,S=p}$ for $g=0,1$ follows from the above two relations as

$$Y^{(1)}_{G=1,S=p} = \frac{(S^{(a)}+S^{(p)})Y^{(1)}_{G=1,E=1} - S^{(a)}Y^{(1)}_{S=a}}{S^{(p)}} \qquad \text{(A.4)}$$

$$Y^{(1)}_{G=0,S=p} = \frac{(S^{(n)}+S^{(p)})Y^{(1)}_{G=0,E=0} - S^{(n)}Y^{(1)}_{S=n}}{S^{(p)}} \qquad \text{(A.5)}$$

By using (A.1) and the left sides of (A.2) and (A.3) we thus obtain (A.4)-(A.5) in terms of the cohort data distribution. Thus the above is the inverse mapping under Assumptions 1-3, that maps the cohort joint distribution of G, E, and Y back to the distribution $(P(G),P(S),P(Y(g) \mid S))$ of interest.

Note that there are distributions $P(G,E,Y)$ that cannot arise from any distribution $(P(G), P(S), P(Y(g) \mid S))$ with Assumptions 1-3. For such a distribution $P(G,E,Y)$, the falsity of Assumptions 1-3 is verifiable because the resulting values of the above procedure to such a distribution $P(G,E,Y)$ are outside the probability space (0,1).

5 Appendix B: Consequences of the combined case-control instrumental variables design

For simplicity assume we are within levels of the matching covariates X. The target quantities in the form of the joint distribution $(P(G), P(S), P(Y(g) \mid S))$ can be calculated using case-control data along with the external prevalence. Let $P^{\text{case}}_{\text{control}}$ denote the distribution in the nested case-control study. Consider first calculating the probabilities $S^{(n)}$, $S^{(a)}$, and $S^{(p)}$:

$$
\begin{aligned}
S^{(n)} &= P(E=0 \mid G=1) \text{ (from A.1 of Appendix A)} \\
&= \frac{\sum_{y'} P(G=1,E=0|Y=y')P(Y=y')}{\sum_{y',e'} P(G=1,E=e'|Y=y')P(Y=y')} \\
&= \frac{\sum_{y'} P^{\text{case}}_{\text{control}}(G=1,E=0,Y=y') \cdot \frac{P(Y=y')}{P^{\text{case}}_{\text{control}}(Y=y')}}{\sum_{y',e'} P^{\text{case}}_{\text{control}}(G=1,E=e',Y=y') \cdot \frac{P(Y=y')}{P^{\text{case}}_{\text{control}}(Y=y')}} \\
&= \frac{\sum_{y'} Y^{\text{case}}_{\text{control}}(y')_{G=1,E=0} \cdot E^{\text{case}}_{\text{control}}(0)_{G=1} \cdot \frac{P(Y=y')}{P^{\text{case}}_{\text{control}}(Y=y')}}{\sum_{y',e'} Y^{\text{case}}_{\text{control}}(y')_{G=1,E=e'} \cdot E^{\text{case}}_{\text{control}}(e')_{G=1} \cdot \frac{P(Y=y')}{P^{\text{case}}_{\text{control}}(Y=y')}}, \\
S^{(a)} &= P(E=1 \mid G=0) \text{ (also from A.1 of Appendix A)} \\
&= \frac{\sum_{y'} P(G=0,E=1|Y=y')P(Y=y')}{\sum_{y',e'} P(G=0,E=e'|Y=y')P(Y=y')} \\
&= \frac{\sum_{y'} Y^{\text{case}}_{\text{control}}(y')_{G=0,E=1} \cdot E^{\text{case}}_{\text{control}}(1)_{G=0} \cdot \frac{P(Y=y')}{P^{\text{case}}_{\text{control}}(Y=y')}}{\sum_{y',e'} Y^{\text{case}}_{\text{control}}(y')_{G=0,E=e'} \cdot E^{\text{case}}_{\text{control}}(e')_{G=0} \cdot \frac{P(Y=y')}{P^{\text{case}}_{\text{control}}(Y=y')}}, \qquad \text{and } S^{(p)} = 1 - S^{(a)} - S^{(n)}
\end{aligned}
$$

Now, we can rewrite (A.4) to have that:

$$
Y^{(1)}_{G=0,S=p} = \left\{ \frac{P(G=0,E=0|Y=y)P(Y=y)}{\sum_{y',e'} P(G=0,E=e'|Y=y')P(Y=y')} - \frac{P(G=1,E=0|Y=y)P(Y=y)}{\sum_{y',e'} P(G=1,E=e'|Y=y')P(Y=y')} \right\} / S^{(p)} \tag{A.6}
$$

$$= \left\{ \frac{Y^{\text{case}}_{\text{control}}(1)_{G=0,E=0} \cdot E^{\text{case}}_{\text{control}}(0)_{G=0} \cdot \frac{P(Y=y)}{P^{\text{case}}_{\text{control}}(Y=y)}}{\sum_{y',e'} Y^{\text{case}}_{\text{control}}(y')_{G=0,E=e'} \cdot E^{\text{case}}_{\text{control}}(e')_{G=0} \cdot \frac{P(Y=y')}{P^{\text{case}}_{\text{control}}(Y=y')}} \right.$$
$$\left. - \frac{Y^{\text{case}}_{\text{control}}(1)_{G=1,E=0} \cdot E^{\text{case}}_{\text{control}}(0)_{G=1} \cdot \frac{P(Y=y)}{P^{\text{case}}_{\text{control}}(Y=y)}}{\sum_{y',e'} Y^{\text{case}}_{\text{control}}(y')_{G=1,E=e'} \cdot E^{\text{case}}_{\text{control}}(e')_{G=1} \cdot \frac{P(Y=y')}{P^{\text{case}}_{\text{control}}(Y=y')}} \right\} / S^{(p)}, \tag{A.7}$$

and

$$Y^{(1)}_{G=1,S=p} = \left\{ \frac{P(G=1,E=1|Y=y)P(Y=y)}{\sum_{y',e'} P(G=1,E=e'|Y=y')P(Y=y')} \right.$$
$$\left. - \frac{P(G=0,E=1|Y=y)P(Y=y)}{\sum_{y',e'} Pr(G=0,E=e'|Y=y')P(Y=y')} \right\} / S^{(p)}$$
$$= \left\{ \frac{Y^{\text{case}}_{\text{control}}(1)_{G=1,E=1} \cdot E^{\text{case}}_{\text{control}}(1)_{G=1} \cdot \frac{P(Y=y)}{P^{\text{case}}_{\text{control}}(Y=y)}}{\sum_{y',e'} Y^{\text{case}}_{\text{control}}(y')_{G=1,E=e'} \cdot E^{\text{case}}_{\text{control}}(e')_{G=1} \cdot \frac{P(Y=y')}{P^{\text{case}}_{\text{control}}(Y=y')}} \right.$$
$$\left. - \frac{Y^{\text{case}}_{\text{control}}(1)_{G=0,E=1} \cdot E^{\text{case}}_{\text{control}}(1)_{G=0} \cdot \frac{P(Y=y)}{P^{\text{case}}_{\text{control}}(Y=y)}}{\sum_{y',e'} Y^{\text{case}}_{\text{control}}(y')_{G=0,E=e'} \cdot E^{\text{case}}_{\text{control}}(e')_{G=0} \cdot \frac{P(Y=y')}{P^{\text{case}}_{\text{control}}(Y=y')}} \right\} / S^{(p)}$$

where $Y^{\text{case}}_{\text{control}}(y')_{G=g',E=e'}$ and $E^{\text{case}}_{\text{control}}(e')_{G=g'}$ denote the induced distributions $P^{\text{case}}_{\text{control}}(Y=1|G=g,E=e)$ and $P^{\text{case}}_{\text{control}}(E=e|G=g)$ in the case-control population. Note that the above forms generally differ from those that result from the naive application of (A.4) and (A.5) to case-control data when $P(Y)$ is not equal to $P^{\text{case}}_{\text{control}}(Y)$. This also serves as a constructive proof that when using instrumental variables the case-control design must be addressed.

REFERENCES

[1] Mauro, V. Statistical methods for the evaluation of policies for firms. PhD Thesis (first draft 2007). University of Florence, Department of Statistics.

[2] Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* 2007; **16**(4):309-330.

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

[4] Davey Smith G, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology* 2004; **33**(1):30-42

[5] Davey Smith G, Ebrahim S, Lewis S, Hansell AL, Palmer LJ, Burton PR. Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* 2005; **366**(9495):1484-98

[6] Davey Smith G. Randomised by (your) god: robust inference from an observational study design. *Journal of Epidemiology and Community Health*. 2006. **60**:382-388.

[7] Davey Smith G, Timpson N, Ebrahim S. Strengthening causal inference in cardiovascular epidemiology through Mendelian randomization. *Annals of Medicine* 2008; **40**:524-541.

[8] Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* 2008; **27**:1133-1163.

[9] Angrist J, Imbens J, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996; **91**:444-455.

[10] Erlinger T, Platz E, Rifai N, Helzlsouer K. C-reactive protein and the risk of incident colorectal cancer. *Journal of the American Medical Association* 2004; **291**:585-590.

[11] Prentice R, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika* 1979; **66**:403-411.

[12] Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. *Annual Review Public Health* 1998; **19** 17-34.

[13] Shetty KD, Vogt WB, Bhattacharya J. Hormone replacement therapy and cardiovascular health in the United States. *Medical Care* 2009; **47** 600-606.

[14] Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006; **4**: 360-372.

[15] Vytlacil, E. Independence, monotonicity, and latent index models: an equivalence result. *Econometrica* 2002; **70**(1):331-341.

[16] Pearl, J. Causality: models, reasoning, and inference. *Cambridge University Press:* Cambridge, 2000.

[17] Tsilidis K, Helzlsouer K, Smith MW, et al. Association of common polymorphisms in il10, and in other genes related to inflammatory response and obesity with colorectal cancer. *Cancer Causes and Control* 2009; **20**:1739-1751.

[18] Allin KH, Bojesen SE, and Nordestgaard BG. Baseline C-reactive protein is associated with incident cancer and survival in patients with cancer. *Journal of Clinical Oncology* 2009; **27**:2217-2224.

[19] Bowden J and Vansteelandt S. Mendelian randomization analysis of case-control data using structural mean models. *Statistics in Medicine* 2011; **30**: 678-694.

[20] Shinohara RT, Frangakis CE, Platz EA, and Tsilidis K. Estimating effects by combining instrumental variables with case-control designs: the role of principal stratification. Johns Hopkins University Department of Biostatistics Working Papers no. 198, 2009.

[21] Timpson N, Lawlor D, Harbord R, et al. C-reactive protein and its role in metabolic syndrome: mendelian randomisation study. *Lancet* 2005; **366**:1954-1959.

[22] Surveillance Epidemiology and End Results [database online]. Bethesda, MD: National Cancer Institute, 2009.

[23] Dehghan A, Kttgen A, Yang Q, Hwang SJ, Kao WL, Rivadeneira F, Boerwinkle E, Levy D, Hofman A, Astor BC, Benjamin EJ, van Duijn CM, Witteman JC, Coresh J, Fox CS. Association of three genetic loci with uric acid concentration and risk of gout: a genome-wide association study. *Lancet* 2008; **372** 1953–1961.

[24] Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K. A high-resolution recombination map of the human genome. *Nature Genetics* 2002; **31**:241-247.

[25] Joffe, M. Principal stratification and attribution prohibition: good ideas taken too far. *The International Journal of Biostatistics* 2011; **7**(1):35.

[26] Li F, Frangakis C. Polydesigns in causal inference. *Biometrics* 2006; **62**:343-351.

[27] Constantinou P. *Mendelian randomisation.* Master's thesis; University of Cambridge, 2009

[28] Rubin D. B. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* 1974; **66**:688-701.

[29] Rubin, D.B. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics.* 1978; **6**: 34-58.

[30] Robins, J. M., Greenland, S., and Hu, F.-C. Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome. *Journal of the American Statistical Association* (with discussion) 1999; **94**: 687-712.

[31] Frangakis C, Brookmeyer R, Varadhan R, Mahboobeh S, Vlahov D, Strathdee S. Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a needle exchange program. *Journal of the American Statistical Association* 2004; **99**:239-249.

[32] Heckman JJ, Urzua S and Vytlacil E. Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics* 2006; **88**:389-342.

[33] Vansteelandt S and Goetghebeur E. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Sociey: Series B* 2003; **65**:817-835.

[34] Cole S, Hernan M. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* 2008; **168**:656-664.

[35] Abadie A. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics* 2003; **113**:231-263.

[36] Tan Z. Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association* 2006; **101**:1607-1618.

[37] Imbens G, Rubin D. Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* 1997; **25**:305-327.