Obtaining causal estimates of therapeutic effects in observational studies:

the usefulness and validity of
physician's preference as an instrumental variable

Anna G.C. Boef

Obtaining causal estimates of therapeutic effects in observational studies:

the usefulness and validity of physician's preference as an instrumental variable.

A.G.C. Boef

ISBN: 978-94-90858-44-5
Layout: Drukkerij Mostert en Van Onderen, Leiden.
Published by: Drukkerij Mostert en Van Onderen, Leiden.

The work described in this thesis was performed at the Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands.
Studies published in this thesis were supported by a grant from the Netherlands Organisation for Health Research and Development (ZonMw, grant number 152002040).

Obtaining causal estimates of therapeutic effects in observational studies:

the usefulness and validity of
physician's preference as an instrumental variable.

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 10 februari 2016
klokke 11.15 uur

door

Anna Gunnel Christina Boef
geboren te Zoeterwoude
in 1988

Promotor:	Prof. dr. J.P. Vandenbroucke
Copromotores:	Dr. S. le Cessie
	Dr. O.M. Dekkers
Leden promotiecommissie:	Prof. dr. T. Stijnen
	Prof. dr. A. de Boer (Universiteit Utrecht)
	Prof. dr. M.E. Numans
	Prof. dr. D.A. Lawlor (University of Bristol)

Table of Contents

Chapter 1

General introduction

1

Introduction

In this thesis we investigate the usefulness and validity of instrumental variable analysis in clinical epidemiological studies, particularly using physician's preference as the instrument. First applications of this method in clinical epidemiology were met with enthusiasm, as it showed promise as a method to handle confounding by indication in a manner mimicking randomisation in a randomised controlled trial. We aim to evaluate this method in both practical applications using existing data and in simulation studies, to expose potential problems and limitations of the method and to identify the settings and types of questions for which it is most useful.

Confounding by indication in observational studies

For many important medical questions and diseases there are no data from randomised controlled trials (RCT) to guide medical practice. Moreover, even if RCTs exist, they might be of limited value for clinical practice if the study insufficiently addressed clinically relevant endpoints or had an insufficient follow-up period. A second problem is that effects of clinical trials often have limited external validity because of highly selected study populations and the controlled set-up of the study. There is therefore an urgent need for methods that enable valid estimation of therapeutic effects from observational studies. However, observational data analyses of anticipated effects of therapy are always suspected to be strongly confounded by factors that determine prognosis, because patient characteristics related to the patient's prognosis also influence the decision how to treat the patient. Technically this is called 'confounding by indication'.[1] Usual methods for dealing with confounding, such as stratification, matching or multivariable analysis can adjust for measured factors only and therefore rely on the often implausible assumption of no unmeasured confounding.[2;3] This also holds for techniques using propensity scores[4] or confounder scores.[2;5]

Instrumental variable analysis

A potential solution to the problem of unmeasured confounding is instrumental variable analysis, a technique which originates in econometrics. The first record of its use in econometrics to bypass confounding is in an appendix of *The tariff on animal and vegetable oils* by Philip Wright, published in 1928. This appendix describes how "the introduction of additional factors" can be used to estimate elasticity of supply and demand: factors which affect demand conditions without affecting cost conditions can be used to estimate elasticity of supply and factors which affect cost conditions without affecting demand conditions can be used to estimate elasticity of demand.[6;7] The term "instrumental variable" was first used in 1945 by Olav Reiersøl.[8;9]

Instrumental variable analysis has flourished in econometrics, in which it is a main tool. However, in most of this thesis we will concentrate on the recent methodologic literature about instrumental variables in epidemiology, because this literature takes into account the developments within econometrics and additionally covers issues and considerations specific to instrumental variables in epidemiology.

The general idea of instrumental variable analysis is to estimate the effect of the exposure on the outcome by utilising a factor (the instrument or instrumental variable) which is related to the exposure, but which is not related to the outcome other than through its association with the exposure (which we will define more formally later).[10] Any difference in the outcome between levels of the instrument can then be attributed to the difference in the exposure between the levels of the instrument. An early application in the estimation of effects of medical treatment was a 1994 study by McClellan et al, which investigated the effect of intensive treatment of myocardial infarction patients on long-term survival. Differential distance to a catheterisation or revascularisation hospital versus a hospital without these facilities was used as an instrumental variable.[11] The use of instrumental variables and specifically the use of clinician preference as an instrumental variable for the estimation of treatment effects, was discussed in a 1998 paper in a statistical journal, using an example of a study in orthodontics which we will consider in further detail later.[12] Introductory papers on instrumental variable analysis subsequently appeared in 1998 in a public health journal[13] and in 2000 in an epidemiological journal.[14] The subsequent decade saw a great increase in both applications of instrumental variable analysis for estimation of effects of therapy[15-19] and theoretical papers on instrumental variable analysis in the field of epidemiology.[3;10;20-22]

Using physician's preference as an instrumental variable

Treatment choices by medical doctors are based on a mix of prognostic characteristics of the patient and an overall preference for a certain type of therapy. If two physicians differ in their overall preference for a certain type of therapy, they may make different treatment decisions when presented with identical patients. Physicians' preferences can therefore result in variation in treatment which is unrelated to patient characteristics and prognosis and can therefore be used as an instrument in an instrumental variable analysis. The principle is that differences in outcomes between similar groups of patients treated by physicians with different treatment preferences are ascribed to the differences in treatment prescription (that occur solely due to the differences in preference). Of course, differences in physician's preference will usually not result in different treatments for all patients as for some patients physicians are likely to choose the same treatment, e.g. for overriding prognostic reasons.

Physician's preference is not a directly measurable characteristic and therefore needs to be estimated from the data in some way. An estimate of the physician's preference at the time of treating a given patient is usually obtained from the treatment choice by the physician for (one or more) previous patients (e.g., previous prescriptions). A patient is then analysed according to his probability of treatment based on his physician's treatment preference as indicated by previous prescriptions, instead of according to his own actual treatment. The prognostic characteristics of previous patients have no bearing on the present patient (given that the main instrumental variable assumptions which will be discussed later hold). An indicator of therapy based on previous patients therefore should not be related to the baseline prognosis of the present patient.

The idea of using clinician's treatment preference as a treatment assignment instrument was proposed by Korn and Baumrind in the setting of a study investigating the effect of tooth extraction on numerous physical measurements in orthodontic patients with crowding and irregularities of their teeth and jaws. Several different randomised and observational study designs were proposed. In brief, the most relevant proposal a design in which the orthodontists who treated the patients would evaluate blinded pre-treatment records of each other's patients, deciding whether they would have treated these patient with extraction. The analysis would then be restricted to those patients on whose treatment the orthodontists disagree. As an alternative to this design, which closely resembles an instrumental variable approach, a true instrumental variable analysis using orthodontist's preference as an instrument was proposed (although not carried out in this paper).[12] Several years later, physician's prescribing preference was first formally used as an instrumental variable in a pharmaco-epidemiological study investigating the effect of selective COX-2 inhibitors in comparison to non-selective non-steroidal anti-inflammatory drugs (NSAIDs) on gastrointestinal complications. The previous prescription of the prescribing physician (for a COX-2 inhibitor or non-selective NSAID) was used as a proxy for the preference of the prescribing physician in an instrumental variable analysis.[15] Numerous other studies have used physician's preference as an instrumental variable since.[17;19;23]

Main assumptions for instrumental variable analysis

In order to be valid, an instrumental variable should fulfil three main assumptions. In case of physician's preference, these assumptions are as follows:[3;15;22;24]

1. Variation in physician's preference is related to the probability of treatment.
2. Physician's preference does not affect the outcome in other ways than through treatment choice (exclusion restriction).
3. Physician's preference does not share causes with the outcome, e.g. is not related to characteristics of a physician's patient population (independence assumption).

Figure 1 depicts these assumptions in a directed acyclic graph: physician's preference Z affects the probability of receiving treatment X. An association of Z and X would also be sufficient for assumption 1. Assumption 2 is represented by the absence of an arrow from Z to Y, i.e. the absence of a direct effect of physician's preference on the outcome or an effect of physician's preference on outcome through any other factor than treatment X. If physician's preference for treatment X also affected prescription of other treatments which affect outcome Y, this assumption would be violated. Assumption 3 is represented by the absence of an arrow from the patient characteristics C and U to physician's preference Z and by the absence of any other common causes of Z and Y. This assumption would be violated if physician's preference were related to characteristics of the physician's patient population. If physician's preference were related to measured patient characteristics only (i.e. an arrow from C to Z) this could be resolved by correcting for these measured patient characteristics in the instrumental variable analysis.

The above assumptions are only sufficient for the estimation of the upper and lower bounds of the average treatment effect.[22;24] An additional assumption is required to obtain a point estimate and the interpretation of the point estimate depends on this additional assumption, as will be discussed in more detail later.

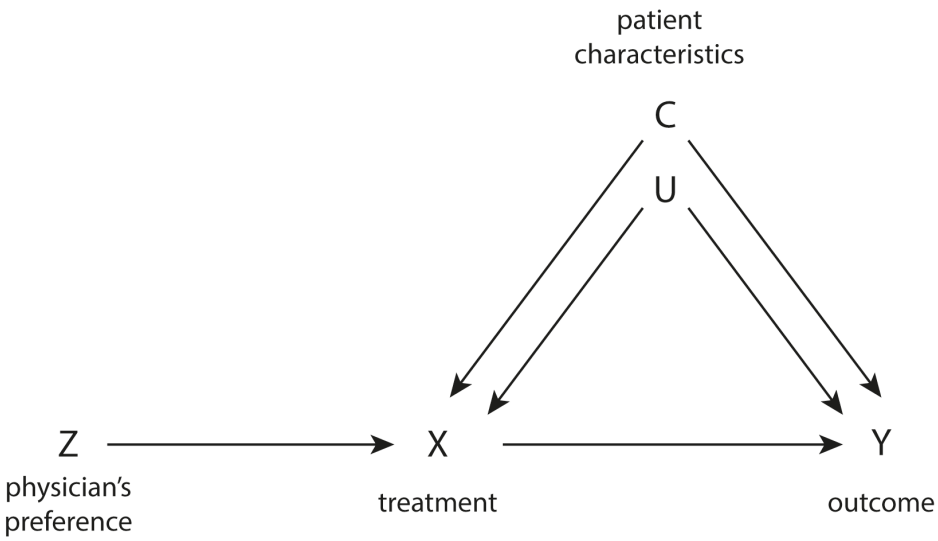


Figure 1. *Directed acyclic graph of the three main assumptions underlying the use of physician's preference as an instrumental variable. C denotes measured confounders of the X-Y relation, U denotes unmeasured confounders of the X-Y relation.*

1

Statistical methods for instrumental variable analysis

Numerous statistical methods to estimate the effect of an exposure on the outcome through instrumental variable analysis exist. Two standard methods for estimation of a risk difference or mean difference are the Wald estimator and two-stage least squares regression.

Intuitively, in case of a binary instrument, the Wald estimator divides the difference in the outcome between the two levels of the instrument by the difference in the exposure between the two levels of the instrument. This yields a risk difference (if the outcome is binary) or a mean difference (if the outcome is continuous) between two exposure or treatment options (if the exposure is binary) or for a difference of 1 unit in the exposure (if the exposure is continuous).

More formally, in case of a binary instrument Z the effect of exposure (treatment) X on outcome Y estimated through the Wald estimator is as follows:[21;22]

$$\hat{\beta}_{IV} = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[X|Z=1] - E[X|Z=0]}$$

in which the numerator is the mean difference or risk difference in the outcome between the two levels of the instrument and the denominator is the mean difference or risk difference in the exposure between the two levels of the instrument.

Two-stage least squares instrumental variable regression involves two linear regression stages. The first stage is a regression of the exposure X (treatment) on the instrument Z (and optionally, potential confounders C). This is the used to obtain a predicted treatment . The second stage is a regression of the outcome Y on the predicted treatment (and optionally, potential confounders C).[3] If no confounders are included, the two-stage least squares estimate equals the Wald estimate. The inclusion of potential confounders C in both regression stages (importantly, the same covariates should then be included in both stages) allows adjusting for potential confounding of the instrument-outcome relation.[21] Like the Wald estimate, the two-stage least squares estimate is a mean difference or risk difference in the outcome for a unit difference in the exposure.

This highlights the main limitation of these methods: for binary outcomes the estimate of interest is usually not a risk difference, but an odds ratio or risk ratio. Several papers discuss potential methods for the estimation of odds ratios[25;26] or risk ratios,[22;26] but the various potential methods make different assumptions, estimate different causal parameters in case of odds ratios and their results can vary in practical applications.[26]

An additional assumption for point identification of a treatment effect

The three assumptions mentioned above are only sufficient for the estimation of the bounds on a treatment effect.[22] A further assumption is required to obtain a point estimate. There are several different options for this 'fourth' instrumental variable assumption, some of which will be discussed here. Importantly, the interpretation of the IV effect estimate is different for each of these options, as we will explain below.

Under the assumption that the treatment effects are homogeneous, the estimator as described above is an asymptotically unbiased estimate of the average treatment effect in the population .[22;24] However, this assumption is often unrealistic because effects of treatments likely vary depending on, for example, severity of disease. Hernan and Robins also argue that for binary outcomes, homogeneity of treatment effects is logically impossible (unless treatment has no effect for any subject).[22] When treatment effects are heterogeneous, an alternative fourth assumption is the deterministic monotonicity assumption. For a binary instrument Z and treatment X this means that only three types of patients may exist, 'always takers' (who receive , 'never takers' and 'compliers' (patients who would receive treatment X at if Z=1 and but not at Z=0). There may be no 'defiers': patients who would receive treatment X at if Z=0 and but not at Z=1. The effect estimated under this assumption is a local average treatment effect (LATE), which for a binary instrument and treatment corresponds to the treatment effect among the 'compliers': i.e. among those patients who would receive treatment X at instrument value A but not at instrument value B.[22;24;27] More generally formulated the instrument must affect treatment monotonically in one direction for all subjects.[22;28;29] Hernan and Robins also specifically discuss the deterministic monotonicity assumption for continuous instruments.[22] An alternative version of the monotonicity assumption is the stochastic monotonicity assumption, which states that instrument should be related to treatment monotonically *across* subjects within strata of a sufficient set of measured and unmeasured common causes of treatment and the outcome.[28;29] For a binary instrument this means that within each of these strata, the compliers should outnumber the defiers. The effect estimated under this assumption is a strength of IV weighted average treatment effect: i.e. a weighted average of the effects in these different strata, with more weight given to strata in which the instrument is stronger.[28]

Mendelian randomisation

Epidemiologic studies investigating aetiology rather than therapeutic effects frequently use a specific form of instrumental variable analysis called Mendelian randomisation, which uses genetic variation as the instrument. Like observational studies of therapeutic effects, observational etiologic studies suffer from confounding

due to unmeasured factors. Another issue in etiologic studies and another reason for performing a Mendelian randomisation study is reverse causation: the observed exposure-outcome association can be due to the 'outcome' causing the 'exposure' rather than the other way round. A well-known example is the observation that low serum cholesterol levels are associated with occurrence of cancer: presence of occult or early-stage cancer may lower cholesterol levels rather than low cholesterol levels causing cancer. Katan proposed to investigate whether genetically low cholesterol levels (i.e. apolipoprotein E variants) are also associated with cancer. Because cancer cannot affect apolipoprotein E variants the issue of reverse causation can hereby be avoided.[30] Furthermore, because genetic variants are randomly allocated from parents to offspring, in Mendelian randomisation studies these variants will generally be unrelated to other factors which affect the outcome. Confounding of the association between genetically determined exposure levels and the outcome is therefore not expected.[31]

Aims and outline of this thesis

We start by examining the concept of physician's prescribing preference itself: first and foremost whether it exists - it can always be argued that observed variation in prescription patterns among physicians in epidemiological datasets is due to differences in their patient populations. In **Chapter 2** we therefore use survey data asking general practitioners whether they would treat eight fictitious subclinical hypothyroidism cases to investigate whether differences in prescription patterns are still present when physicians are presented with the same patients. Additionally we investigate the plausibility of the deterministic and stochastic monotonicity assumptions for physician's prescribing preference by examining the preference patterns in these survey data.

Most examples of applications of physician's preference-based instrumental variable analysis were performed in large pharmacoepidemiologic databases.[15;17;19] In **Chapter 3** we aim to investigate how valid and useful physician's preference-based instrumental variable analysis is in clinical epidemiological studies of a more typical size, i.e. several hundred patients. To this aim, we compare instrumental variable estimates of the effect of preoperative corticosteroids on mechanical ventilation time and duration of intensive care and hospital stay, occurrence of infections, atrial fibrillation, heart failure and delirium using routine care data of elective cardiac surgery patients to estimates from conventional analyses in the same data and to results of a recent randomised controlled trial.

After this application, we use theoretical derivations and simulations in **Chapter 4** to investigate how sample size influences how instrumental variable analyses and

conventional analyses compare with regard to the average deviation of the estimates from the true effect, depending on the strength of the instrument and the level of confounding.

In a situation in which no confounding by (contra-)indication is expected, e.g. if the outcome is an unknown or unpredictable side-effect, instrumental variable analysis should give the same estimate as a conventional analysis. We investigate whether this holds for the effect of third generation oral contraceptives versus second generation oral contraceptives on the occurrence of venous thromboembolism in **Chapter 5,** comparing instrumental variable estimates using general practitioner's preference as an instrument to conventional estimates.

Chapter 6 is a comment to a publication on the reporting of instrumental variable analyses, in which we provide a suggestion for an additional reporting step: presenting the outcome by strata of the instrument prior to performing a formal instrumental variable analysis. This has an interest in itself since the assumptions required are less stringent than for the formal instrumental variable analysis.

In the last two chapters we move from the use of physician's preference as an instrumental variable for estimating effects of therapy to the use of genetic variation as an instrumental variable in etiologic studies. In **Chapter 7** we provide a meta-epidemiological overview of the different methodological approaches used in Mendelian randomisation studies and evaluate the reporting of the statistical methods and the discussion of the plausibility of Mendelian randomisation assumptions. In **Chapter 8** we explain why selection bias may exist if Mendelian randomisation studies are performed in elderly populations.

1

References

(1) Greenland S, Neutra R. Control of confounding in the assessment of medical technology. *Int J Epidemiol* 1980;9:361-367.

(2) Klungel OH, Martens EP, Psaty BM et al. Methods to assess intended effects of drug treatment in observational studies are reviewed. *J Clin Epidemiol* 2004;57:1223-1231.

(3) Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. *J Clin Epidemiol* 2009;62:1226-1232.

(4) Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41-55.

(5) Miettinen OS. Stratification by a multivariate confounder score. *Am J Epidemiol* 1976;104:609-620.

(6) Wright PG. Appendix B. *The Tariff on Animal and Vegetable Oils*. New York: Macmillan; 1928.

(7) Stock JH, Trebbi F. Who Invented Instrumental Variable Regression? *Journal of Economic Perspectives* 2003;17:177-194.

(8) Reiersøl O. Confluence Analysis by Means of Instrumental Sets of Variables. *Arkiv för Matematik, Astronomi och Fysik* 2015;32a:1-1196.

(9) Angrist JD, Krueger AB. Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives* 2001;15:69-85.

(10) Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *Int J Biostat* 2007;3:Article.

(11) McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA* 1994;272:859-866.

(12) Korn EL, Baumrind S. Clinician Preferences and the Estimation of Causal Treatment Differences. *Statistical Science* 1998;13:209-235.

(13) Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. *Annu Rev Public Health* 1998;19:17-34.

(14) Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000;29:722-729.

(15) Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* 2006;17:268-275.

(16) Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables II: instrumental variable application-in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. *J Clin Epidemiol* 2009;62:1233-1241.

(17) Schneeweiss S, Seeger JD, Landon J, Walker AM. Aprotinin during coronary-artery bypass grafting and risk of death. *N Engl J Med* 2008;358:771-783.

(18) Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA* 2007;297:278-285.
(19) Wang PS, Schneeweiss S, Avorn J et al. Risk of death in elderly users of conventional vs. atypical antipsychotic medications. *N Engl J Med* 2005;353:2335-2341.
(20) Martens EP, Pestman WR, de BA, Belitser SV, Klungel OH. Instrumental variables: application and limitations. *Epidemiology* 2006;17:260-267.
(21) Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf* 2010;19:537-554.
(22) Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006;17:360-372.
(23) Davies NM, Smith GD, Windmeijer F, Martin RM. COX-2 selective nonsteroidal anti-inflammatory drugs and risk of gastrointestinal tract complications and myocardial infarction: an instrumental variable analysis. *Epidemiology* 2013;24:352-362.
(24) Swanson SA, Hernan MA. Commentary: how to report instrumental variable analyses (suggestions welcome). *Epidemiology* 2013;24:370-374.
(25) Rassen JA, Schneeweiss S, Glynn RJ, Mittleman MA, Brookhart MA. Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. *Am J Epidemiol* 2009;169:273-284.
(26) Palmer TM, Sterne JA, Harbord RM et al. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. *Am J Epidemiol* 2011;173:1392-1403.
(27) Imbens GW, Angrist JD. Identification and Estimation of Local Average Treatment Effects. *Econometrica* 1994;62:467-475.
(28) Small DS, Tan Z, Lorch SA, Brookhart MA. Instrumental variable estimation when compliance is not deterministic: the stochastic monotonicity assumption. 2014.
(29) Small DS, Tan Z. A stochastic monotonicity assumption for the instrumental variables method. 2007.
(30) Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet* 1986;1:507-508.
(31) Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* 2008;27:1133-1163.

Chapter 2

Physician's prescribing preference as an instrumental variable: exploring assumptions using survey data.

Anna G.C. Boef, Saskia le Cessie, Olaf M. Dekkers, Peter Frey, Patricia M. Kearney, Ngaire Kerse, Christian D. Mallen, Vera J.C. McCarthy, Simon P. Mooijaart, Christiane Muth, Nicolas Rodondi, Thomas Rosemann, Audrey Russell, Henk Schers, Vanessa Virgini, Margot W.M. de Waal, Alex Warner, Jacobijn Gussekloo, Wendy P.J. den Elzen

Epidemiology 2016 (accepted for publication)

Abstract

Background: Physician's prescribing preference is increasingly used as an instrumental variable (IV) in studies of therapeutic effects. However, differences in prescribing patterns among physicians may reflect different preferences or differences in case-mix. Furthermore, there is debate regarding the plausibility of possible assumptions required for obtaining a point estimate using physician's preference as an instrument.

Methods: A survey containing eight fictitious cases of women with subclinical hypothyroidism was sent to general practitioners (GPs) in The Netherlands, the United Kingdom, New Zealand, Ireland, Switzerland and Germany. GPs were asked whether they would prescribe levothyroxine to these cases. We investigated (1) whether variation in physician's preference was observable and to what extent it was explained by characteristics of the GPs and their patient populations and (2) to what extent the data were compatible with deterministic and stochastic monotonicity assumptions.

Results: There was substantial variation in levothyroxine prescriptions amongst the 526 responding GPs. Between-GP variance in levothyroxine prescriptions (on a logit scale) was 9.89 (95% CI 8.02;12.20) in the initial mixed-effects logistic model, 8.26 (6.67;10.23) after adding a fixed effect for country and 8.01 (6.47;9.93) after adding GP characteristics. The deterministic monotonicity assumption was falsified by the occurring prescription patterns. For all cases in all countries, the probability of receiving levothyroxine was higher if a different case of the same GP received levothyroxine, which is compatible with the stochastic monotonicity assumption. The data were not compatible with this assumption for a different definition of the instrument.

Conclusions: Our study supports the existence of physician's preference as a determinant in treatment decisions. The deterministic monotonicity assumption will generally not be plausible for physician's preference as an instrument. Depending on the definition of the instrument, the stochastic monotonicity assumption may be plausible.

Introduction

Instrumental variable (IV) analysis is increasingly used in observational studies of therapeutic effects, with the aim of circumventing confounding by indication. This method requires a variable (the instrument) that meets the following conditions: (1) is associated with treatment, (2) does not affect the outcome other than through treatment (exclusion restriction) and (3) does not share a common cause with the outcome (independence assumption).[1;2] One such instrument is physician's prescribing preference, which exploits the notion that prescribing by a medical doctor is influenced not only by prognostic characteristics of the patient, but also by a general preference of the doctor for some type of therapy when different treatment options are available.

Because underlying preference cannot be observed, physician's preference-based IV studies use an estimate of physician's preference based on prescribing behaviour. The question remains, however, whether differences in prescribing behaviour between physicians truly reflect differences in preference rather than just differences in their patient populations. Furthermore, the three main IV conditions described above are only sufficient for the estimation of bounds of a treatment effect.[3] To obtain a point estimate an additional (fourth) assumption is required. The assumption of no heterogeneity of treatment effects, under which the average treatment effect in the population can be estimated, is often implausible.[3] A frequently used alternative is the monotonicity assumption, first described by Imbens and Angrist.[4] According to the original (deterministic) monotonicity assumption, the instrument may only be related to treatment monotonically in one direction for all subjects.[2;4-7] A less strict, stochastic version of the monotonicity assumption has been proposed, as we will explain later.[5-7]

The notion that physician's underlying prescribing preference affects prescribing behaviour cannot be proven in IV study data (at the most, the assumption that physician's estimated prescribing preference is unrelated to characteristics of the physician's patient population can be explored to some extent). Furthermore, the deterministic monotonicity assumption is generally not verifiable within IV study data and the validity of the stochastic monotonicity assumption can only be explored to some extent. Swanson et al recently proposed using a study design in the form of a survey, asking physicians what their treatment decision would be for the same set of cases, to assess the monotonicity assumption empirically.[2] Here we perform such a study, using data from a survey originally performed with the aim of establishing differences in treatment strategies of general practitioners (GPs) for subclinical hypothyroidism by country and by patient characteristics.[8] These data were therefore not primarily

intended for our current study, but can nevertheless provide a valuable insight into the plausibility of the different monotonicity assumptions. Our aims are twofold, (1) to establish whether variation in physician's preference regarding treatment of subclinical hypothyroidism is observable when GPs are presented with the same set of patients and to what extent this variation is explained by characteristics of the GPs and (2) to establish to what extent the data are compatible with the deterministic and stochastic monotonicity assumptions.

Methods

Study data

The survey procedures have been described in detail elsewhere.[8] An online survey was e-mailed to 2710 GPs in The Netherlands, Germany, England, Ireland, Switzerland and New Zealand. It contained eight fictitious cases of women with subclinical hypothyroidism. All cases had a normal BMI, non-specific complaints resulting in fatigue and a normal free thyroxine level. Cases varied in age (70 years/ 85 years), vitality status (vita /vulnerable) and thyroid stimulating hormone (TSH) (6 mU/L/15 mU/L), (Table 1). For each case, GPs were asked if they would start treatment, and, if so, what levothyroxine starting dose they would choose. For the purposes of this study, we only consider the responses on whether treatment would be started. Furthermore, GPs were asked questions about their gender, years of experience as a GP, the percentage of elderly patients registered in their practice, the time since the last diagnosis of subclinical hypothyroidism and the time since last starting levothyroxine treatment in a patient with subclinical hypothyroidism. For the full survey, we refer to Appendix 2 of Den Elzen et al which reports the study for which the survey was originally performed.[8]

Table 1. *Age, vitality status and thyroid stimulating hormone (TSH) of the eight cases in the survey.*

Case	Age	Vitality status	TSH (mU/L)
1	70	Vital	6
2	70	Vulnerable	6
3	70	Vital	15
4	70	Vulnerable	15
5	85	Vital	6
6	85	Vulnerable	6
7	85	Vital	15
8	85	Vulnerable	15

Adapted from Den Elzen et al, British Journal of General Practice 2015.

Possible assumptions for point estimation

Deterministic monotonicity

For a dichotomous instrument the deterministic monotonicity assumption is usually defined as the absence of 'defiers'.[1;2;6;9] The IV analysis then estimates a local average treatment effect (LATE) among the 'compliers'.[1;4] These 'compliers' or 'marginal patients' are those patients who would receive treatment at the 'encouraging' value of the instrument (e.g. preference for treatment), but not at the 'non-encouraging' value of the instrument (e.g. preference for no treatment).[1;5;9;10] As discussed by Swanson et al and Small et al, for physician's preference as an IV, the compliance class (whether the patient is a complier, defier, always taker or never taker) is generally not well defined.[2;6]

Hernan and Robins have formulated the deterministic monotonicity assumption for physician's preference as a continuous instrument.[3] This would translate to the example of subclinical hypothyroidism as follows: if physician A would treat a certain patient with subclinical hypothyroidism with levothyroxine, then all physicians with a preference greater than or equal to the preference of physician A should treat that patient with levothyroxine. It is this assumption which we will assess for our survey data. It would correspond to global monotonicity as described by Swanson et al.[2] (Local monotonicity was also described by Swanson et al: for this somewhat more relaxed version of the assumption monotonicity must hold for specific pairs of physicians.[2]) For continuous instruments, the LATE is a weighted average of treatment effects in multiple subgroups of patients (e.g. subgroups of patients who would receive levothyroxine from physicians with a certain preference but not from physicians with a lower preference).[1;3]

Stochastic monotonicity

The alternative proposed is the stochastic monotonicity assumption, which states that the instrument should be related to treatment monotonically across subjects within strata of a sufficient set of measured and unmeasured common causes of treatment and the outcome.[6]

If we view the cases in our survey not as individual cases but as strata of patients with the same relevant characteristics, the stochastic monotonicity assumption requires GPs' preference to be related to treatment monotonically in one direction across patients in each of these strata. This means that the probability of levothyroxine treatment for patients treated by GPs with preference A should be at least as high as for patients treated by GPs with a lower preference, within all strata of patients.

Under the stochastic monotonicity assumption the effect estimated is a weighted average of treatment effects in the different strata of patients, with more weight given to those strata in which the instrument is strongest.[5;7] Small et al have named this the strength-of-IV weighted average treatment effect (SIVWATE).[6] We point out that in their identification framework for the SIVWATE and LATE, Small et al formulate the three main IV assumptions differently to how we formulated these assumptions in our introduction.[6]

Analysis

Variation in preference for levothyroxine and its determinants

For each GP who completed all survey questions, we calculated the total number of cases treated with levothyroxine, as a measure of the GP's relative preference for treatment with levothyroxine in subclinical hypothyroidism.

To investigate the effect of GP characteristics on their tendency to prescribe levothyroxine, we used mixed-effects logistic regression. All cases completed by the GPs were included, with treatment with levothyroxine (no/yes) as the outcome. We ran the following (pre-specified) models:

Model 1: A random effect for GP and fixed effects for characteristics of the case (age 70 or 85, TSH 6 or 15 mU/L, vital or vulnerable disposition).

Model 2: Model 1 plus a fixed effect for country.

Model 3: Model 2 plus a fixed effect for GP gender and years of experience (<5, 5-10, 11-15, 16-20, 21-25, >25 years).

Model 4: Model 3 plus a fixed effect for percentage of patients in the GP's practice aged ≥65 years (<10%, 10-20%, 20-30%, >30%) and time since last diagnosis of subclinical hypothyroidism (<1 wk, 1 wk-1 mth, 1 mth-1 yr, 1-3 yrs, >3 yrs).

The parameter of interest was the variance of the random effect of the GP ("between-GP variance in preference"), which is calculated on a log odds scale. The interest lies in whether this variance decreases as country and characteristics of the GP are added to the model.

Deterministic monotonicity assumption

To investigate the monotonicity assumption we made a matrix plot[11] for each country, with cases 1 to 8 on the X-axis and the GPs, ordered from highest to lowest preference, on the Y-axis, the colour of each cell indicating whether levothyroxine was prescribed. This was used to visually examine whether the deterministic monotonicity assumption holds. GPs who did not complete the survey were not included in these plots. eFigure 1 shows a matrix plot with the pattern expected if deterministic monotonicity holds completely: physicians with a certain preference always prescribe levothyroxine to those cases for which physicians with the same or a lower preference

prescribe levothyroxine. (From these plots, which show the complete data pattern, it is also possible to derive whether deterministic monotonicity could hold for specific instruments such as treatment of the previous patient of the same GP.)

Stochastic monotonicity assumption

The exact formulation of the stochastic monotonicity assumption depends on the definition of the instrument. Because Small et al discuss the stochastic monotonicity assumption in the context of a binary instrument, using treatment of the previous patient as an example, and because treatment of the previous patient is a frequently used physician's preference-based instrument, we evaluated whether stochastic monotonicity could hold for this instrument. Because all GPs were presented with all cases in the same order, we cannot use the true previous case as instrument. We therefore considered each other case as a potential previous patient, i.e. for each case there were 7 potential previous patients per GP. We denote the potential previous patient as the 'other patient'. Each possible index patient-other patient combination was classified according to the treatment of both patients and summed across GPs to a total per case (per country). For each case we calculated the probability of levothyroxine treatment if the other patient received levothyroxine and if the other patient did not receive levothyroxine.

As a sensitivity analysis we also assessed the stochastic monotonicity assumption for the proportion of all other cases the same GP decided to treat (although we note that Small et al only discussed the stochastic monotonicity assumption with respect to a dichotomous instrument)[6]. We performed this analysis for the two countries with the largest number of responding GPs (The Netherlands and Switzerland).

Missing data

There was a technical problem in the electronic questionnaire sent to the Dutch GPs, resulting in 16 missing answers for case 6. Missing answers due to this technical problem were imputed, using logistic regression (10 imputations) with country, the answers for all other cases and characteristics of the GP as predictors.

Analyses were performed using Stata 12 (College Station, TX: StataCorp LP. 2011).

Results

A total of 526 GPs from 8 countries responded to the survey. eTable 1 lists the response rates per country. The overall response rate was 19% (526/2710) and ranged from 4% (New Zealand) to 41% (The Netherlands). The number of responding GPs ranged from 21 from Ireland to 262 from Switzerland. Table 2 shows the characteristics of the GPs. Of the 526 respondents, 468 (89%) answered all questions and 71% were male. The years of experience ranged from <5 years (8%) to >25 years (29%). Seventy percent of responding GPs had ≥20% patients aged 65 years and over in their practice and the vast majority (91%) had diagnosed a patient with subclinical hypothyroidism within the last year.

Table 2. *Characteristics of participating general practitioners (GPs).*

GP characteristics	n=526
Country	
The Netherlands	129 (25)
United Kingdom	22 (4)
New Zealand	31 (6)
Ireland	21 (4)
Switzerland	262 (50)
Germany	61 (12)
Male	373 (71)
Experience as a GP (years)	
<5	41 (8)
5-10	70 (13)
11-15	90 (17)
16-20	82 (16)
21-25	88 (17)
>25	155 (29)
Patients aged 65 years and over in GP practice (%)	
<10	35 (7)
10-20	122 (23)
20-30	188 (36)
>30	181 (34)
Time since last subclinical hypothyroidism diagnosis	
<1 week	76 (14)
1 week-1 month	194 (37)
1 month-1 year	211 (40)
1-3 years	27 (5)
>3 years	18 (3)

All data are presented as n (%).

Variation in number of levothyroxine prescriptions

Figure 1 displays the distribution per country of the total number of cases for which the GP decided to start levothyroxine. There was substantial variation in this total within each country. The most frequent number of levothyroxine prescriptions was 4 for the UK, New Zealand, Ireland and Switzerland, 0 for The Netherlands and 8 for Germany.

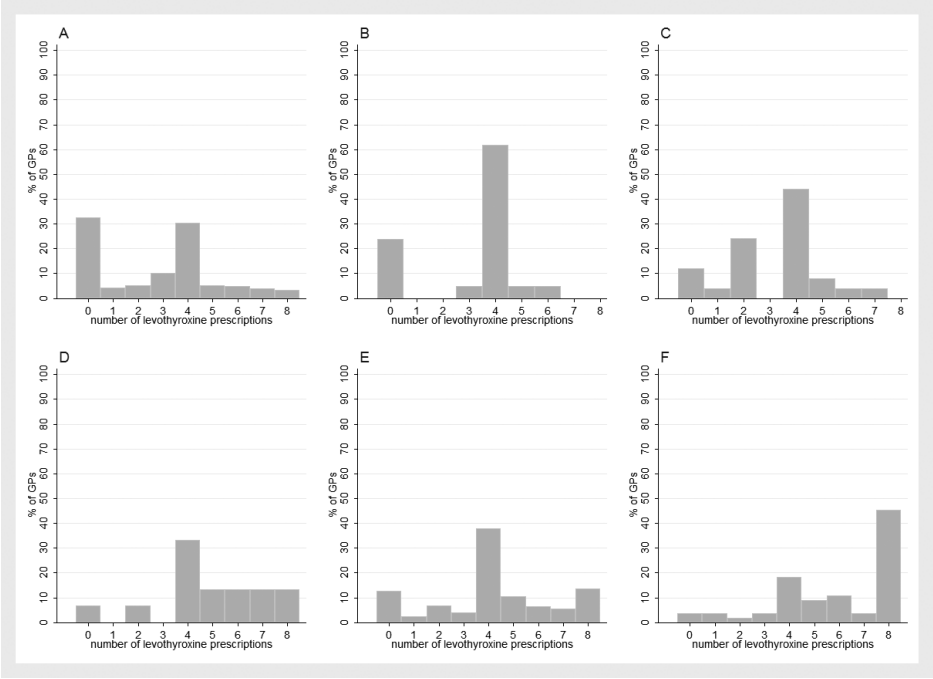


Figure 1. *Distribution per participating country of the number of cases for which a GP would prescribe levothyroxine.*
A. The Netherlands (n=117) *B. United Kingdom (n=21)* *C. New Zealand (n=25)*
D. Ireland (n=15) *E. Switzerland (n=235)* *F. Germany (n=55)*

Association between GP characteristics and treatment preference

Table 3 displays results of the mixed-effects logistic regression used to investigate the effect of GP characteristics on levothyroxine prescription. Country explained some of the variance in levothyroxine prescription between GPs, as shown by the reduction in between-GP variance from 9.91 (95%CI 8.04;12.22) to 8.27 (95%CI 6.68;10.23) after adding a fixed effect for country. Adding GP characteristics (Model 3) resulted in a very small reduction in between-GP variance in treatment to 8.15 (95%CI 6.58;10.10). Adding time since last subclinical hypothyroidism diagnosis and the proportion of patients aged 65 years and over (Model 4) resulted in a similarly small

reduction. There was therefore still substantial variation in levothyroxine prescription among GPs after adjusting for all available patient and doctor characteristics.

Table 3 *Between general practitioner (GP) variance in treatment*

Model	Between GP variance (95% CI)
1: Random effect for GP; fixed effect for age, TSH and vitality status of case	9.89 (8.02;12.20)
2: Model 1 + fixed effect for country	8.26 (6.67;10.23)
3: Model 2 + fixed effect for gender and years of experience	8.15 (6.58;10.10)
4: Model 3 + fixed effect for time since last diagnosis of subclinical hypothyroidism and proportion of patients aged 65 years and over	8.01 (6.47;9.93)

Deterministic monotonicity assumption

Figure 2 shows matrix plots per country of the treatment decisions for each case by each GP. GPs are ordered from highest (8 cases treated) to lowest preference (0 cases treated). The prescription patterns of the UK (Figure 2B) only showed a single violation of deterministic monotonicity: the GP who prescribed levothyroxine to 5 cases treated case 2 while the GP who prescribed levothyroxine to 6 cases did not treat case 2. There were more violations of deterministic monotonicity in the other countries. Treating all cases with a TSH of 15 mU/L was a common pattern in the UK, The Netherlands, New Zealand, Switzerland and Ireland. For example, 75 of 89 GPs who treated 4 cases in Switzerland decided to initiate levothyroxine in cases 3, 4, 7 and 8. In both The Netherlands and Switzerland, most GPs with a lower preference treated (one or more) cases with a high TSH only and most GPs with a higher preference treated at least the high TSH cases. However, there was not a consistent pattern regarding the 5th, 6th or 7th case treated, or the 1st, 2nd or 3rd case treated within those with a TSH of 15 mU/L. Prescribing patterns in Germany differed from those in other countries: many GPs (25 of 55) treated all cases with levothyroxine, and for the other GPs the prescribing patterns were less consistent.

Stochastic monotonicity assumption

Table 4 displays the probability of levothyroxine prescription per case, dependent on treatment of a different patient of the GP. The probability of levothyroxine prescription was higher if the other patient was prescribed levothyroxine for nearly all cases in all countries. Exceptions were case 1 in the UK and in New Zealand, for whom treatment probability did not differ depending on the other patient's treatment. Importantly, there were no cases for whom the probability of levothyroxine was higher if the other

patient did not receive levothyroxine, i.e. the instrument was related to treatment in the same direction for all cases in all countries. The instrument strength (the difference between the probability of the index patient receiving levothyroxine if the other patient received levothyroxine and the probability of the index patient receiving levothyroxine if the other patient did not receive levothyroxine) varied across cases within each country. For example, in The Netherlands, it varied from 20% (case 1) to 47% (case 4).

The sensitivity analysis in which we evaluated the stochastic monotonicity assumption for a continuous instrument (the proportion of all other cases treated) showed violations of this assumption (eTable 2). Although for both countries the probability of treatment increased as the value of the instrument increased for all cases, it did not increase monotonically. Specifically, the probability of treatment was higher if 3/7 other cases were treated than if 4/7 other cases were treated.

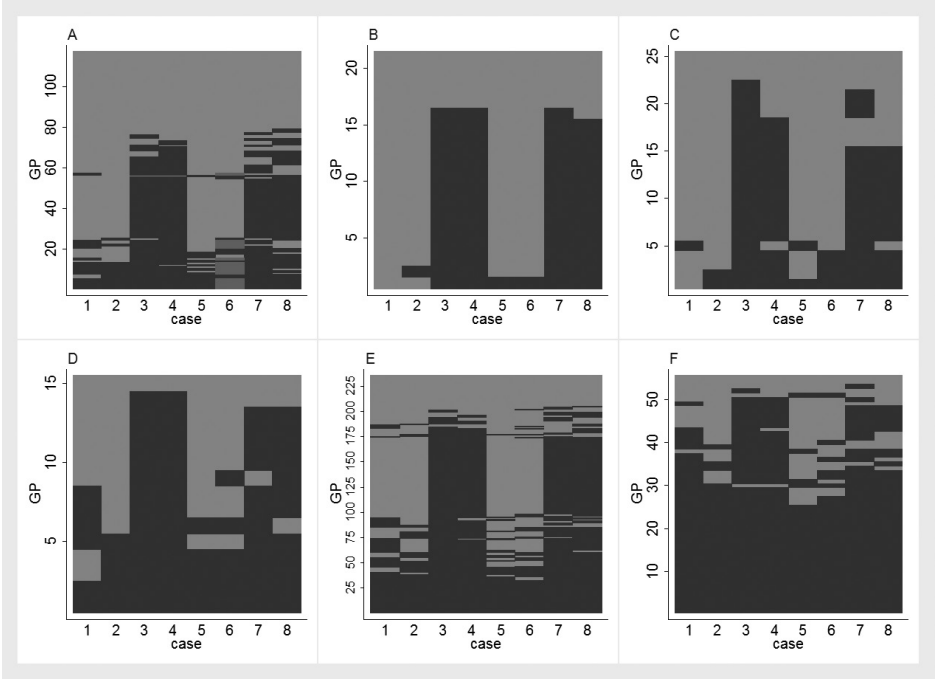


Figure 2. *Matrix plots of the prescription patterns of the GPs within each country. GPs are ordered from highest to lowest preference, with their response for each case indicated by the colour of the cell (yes: dark-grey, no: light-grey, missing: mid-grey). GPs with equal preferences were ordered according to their preferences for case 1 (first yes, then no) through to 8, and subsequently by their identification-number (if all answers were equal).*

A. The Netherlands (n=117) *B. United Kingdom (n=21)* *C. New Zealand (n=25)*
D. Ireland (n=15) *E. Switzerland (n=235)* *F. Germany (n=55)*

Table 4. *Probability of levothyroxine dependent on treatment of a different case by the same general practitioner (GP)*

Case	Country																	
	Netherlands (n=117)			UK (n=21)			New Zealand (n=25)			Ireland (n=15)			Switzerland (n=235)			Germany (n=55)		
	–	+	Δ	–	+	Δ	–	+	Δ	–	+	Δ	–	+	Δ	–	+	Δ
1	7	27	21	0	0	0	4	4	0	27	47	20	17	45	28	50	89	39
2	4	27	23	4	6	2	3	14	11	11	45	34	10	41	31	19	73	54
3	44	90	46	64	100	36	82	100	18	84	100	16	67	96	30	76	97	20
4	45	92	47	64	100	36	60	83	23	84	100	16	62	93	31	68	95	27
5	4	27	22	2	8	5	5	12	6	11	45	34	7	36	29	15	64	50
6	6	26	20	2	8	5	9	25	16	22	50	28	11	38	27	23	72	49
7	39	85	46	64	100	36	63	89	26	65	90	25	61	90	29	65	93	28
8	40	78	38	59	94	35	45	74	29	70	87	17	61	90	30	46	87	41

Percentage of yes answers per case within each country, dependent on the treatment of a different case (the 'other patient') by the same GP. Each other answer of the same GP was used as an 'other patient'. Treatment of the 'previous patient' is indicated by – (no levothyroxine) and + (levothyroxine). The columns indicate the following (in %): – : $Pr[D=1|Z=0]$; +: $Pr[D=1|Z=1]$; Δ: $Pr[D=1|Z=1]$-$Pr[D=1|Z=0]$.

Discussion

This survey study showed marked within-country variation amongst GPs in their tendency to treat patients with subclinical hypothyroidism with levothyroxine. Presenting the same cases to all GPs ensured that observed differences in prescribing behaviour truly reflect differences in preference, rather than differences case-mix. The existence of underlying relative preference for levothyroxine treatment for subclinical hypothyroidism patients amongst GPs as a "pseudo-random" phenomenon is further supported by the very limited decrease in between-GP variance in levothyroxine prescription after adjusting for GP characteristics. Even country explained a relatively small amount of the variation: the within-country variation is considerable compared to between-country differences.

The minimal amount of between-GP variance in levothyroxine prescription explained by GP characteristics within countries is reassuring with regard to main IV assumptions. If GP gender and years of experience were related to relative preference for levothyroxine, this would threaten the validity of the exclusion restriction assumption: years of experience in particular may affect the prognosis of subclinical hypothyroidism patients through other ways than levothyroxine prescription. If the proportion of older

patients were related to preference for levothyroxine this would threaten the validity of the independence assumption: the baseline prognosis of patients would then differ according to GP's preference. With regard to the independence assumption, it is important to make the distinction between physician's preference as assessed in this survey and physician's preference as it is typically used as IV in observational studies. A measure of preference based on previous patients of the physician is typically used: the treatment of these previous patients is determined both by the underlying preference of the physician and by characteristics of these patients.[2] Physicians with the same underlying preference (i.e. who would give the same responses to our survey questions) can have a different case-mix of patients, and an estimate of their preference based on treatment of these patients would then differ. Although the assumption of no confounding seems to hold for underlying preference in our survey data, it may well be violated in observational data if measures of preference based on treatment of previous patients are used, due to confounding by case-mix. This issue of confounding of instruments based on prescribing history was also discussed by Swanson et al.[2]

The preference patterns observed within the six countries deviated in varying degrees from the pattern expected if the deterministic monotonicity assumption would hold. The violation of the deterministic monotonicity assumption in this survey with relatively simple case descriptions indicates it is unlikely to hold for physician's preference as an instrument in true prescription data. For a dichotomous instrument, the bias in the local average treatment effect (LATE) estimate caused by violation of deterministic monotonicity depends on the proportions of compliers and defiers and the difference in treatment effects for compliers and defiers.[9] For a multi-levelled or continuous instrument the bias caused by violation of the deterministic monotonicity assumption will be determined by analogous factors: i.e. the severity and pattern of the deviation from monotonicity, and the level of heterogeneity of treatment effects. In our example, heterogeneity is most likely to exist according to TSH levels, but looking at TSH only, there is relatively little violation of deterministic monotonicity.

In these data, the stochastic monotonicity assumption was not falsified when treatment of a different patient of the same GP was used as an instrument. However, in the sensitivity analysis using the proportion of all other patients of the same GP treated as an instrument, the data were not compatible with the stochastic monotonicity assumption for that instrument. This may be due to the specific setting of the study: a certain proportion of other patients treated often corresponds to a certain pattern of specific cases treated in these data. Overall, these results suggest that the stochastic monotonicity assumption may be plausible for physician's preference-based IV studies, depending on how the instrument is defined. Estimates of preference based on

a larger number of previous patients may be more likely in general to violate stochastic monotonicity, because the probability of treatment must increase monotonically across all levels of these instruments for all strata of patients.

The effect estimate under the stochastic monotonicity assumption is not the LATE but the strength-of-IV weighted treatment effect (SIVWATE), a generalisation of the LATE with a similar interpretation.[6] There has recently been discussion on the usefulness of the LATE. It centres around the question whether the treatment effect for the compliers is a relevant effect,[12;13] particularly because we cannot identify who the compliers are.[12] The SIVWATE has similar drawbacks to the LATE: the interpretation is difficult, since it is a weighted average of effects in strata which we cannot identify and for which we do not know the weights.

The existing survey data used for this study provided a unique opportunity to investigate the assumptions underlying the use of physician's preference as an IV, but also presented some limitations. One limitation is the low response rate, which may have affected our results in various ways. Responding GPs may be more aware of guidelines and more alike in their prescription patterns: i.e. the deterministic monotonicity assumption could be violated to a greater extent in the entire GP population. There may have been more 'random' variation in answers if all GPs had responded (i.e. if underlying preference is a stronger determinant of treatment in the respondents than in GPs overall). This would have reduced the overall strength of GP's preference as an instrument. However, we would not expect it to affect the validity of the stochastic monotonicity assumption for treatment of one other case as the instrument: we do not expect such vastly different patterns among non-respondents that treatment of a particular case would be inversely related to treatment of a different case.

All GPs were presented with the cases in the same order. Random ordering of the cases per GP would have been preferable for assessing preference in the context of an IV. It would have enabled us to use a true 'previous case' for the evaluation of the stochastic monotonicity assumption. Furthermore, the ordering of the cases may have had some influence on answers given for specific cases.

By evaluating the stochastic monotonicity assumptions across these eight patient types (strata) in the survey, we considered the characteristics that define these patient types, i.e. age, vitality status and TSH levels, to be a sufficient set of measured and unmeasured common causes of treatment and the outcome. While this may hold for the simplified survey data, this is unlikely to be a sufficient set in a true patient population. We were therefore only able to evaluate the stochastic monotonicity

assumption for the simplified setting of the survey. Related to this, the fictitious cases in the survey were not intended to represent any particular population of subclinical hypothyroidism patients for whom we may want to estimate the effect of levothyroxine treatment. Rather, the survey was designed in such a manner that characteristics which were thought to be important in the treatment decision varied among the cases. The cases were intended to represent a well-known clinical decision problem: whether to treat subclinical hypothyroidism. In this sense estimating a treatment effect for this group would be of potential interest, although the types of subclinical hypothyroidism patients represented by the cases are limited. For example, the cases were all women and there was no variation in the symptoms with which they presented.

Findings which may be of interest to clinicians are that we can distinguish several groups of factors which are related to the decision whether to treat a patient with subclinical hypothyroidism: characteristics of the patient, country (and its guidelines), and GP's preference. In this setting of treatment of subclinical hypothyroidism, the lack of stringent guidelines leaves substantial room for GP's preference to play a role in treatment decisions. While this would provide an opportunity to utilise this variation in an IV study of the effect of treatment of subclinical hypothyroidism, the ultimate aim of such a study would paradoxically be to reduce this preference-based variation through the development of evidence-based guidelines.

In conclusion, our study supports the existence of physician's preference as a determinant in treatment decisions. Little of the variation in preference was explained by characteristics of the GP or their patient population, indicating that main IV assumptions may be plausible for physician's treatment preferences. The deterministic monotonicity assumption did not hold and will generally not be plausible for physician's preference as an instrument. The stochastic monotonicity assumption may be plausible, depending on how the instrument is defined.

References

(1) Swanson SA, Hernán MA. Commentary: how to report instrumental variable analyses (suggestions welcome). *Epidemiology* 2013;24:370-374.

(2) Swanson SA, Miller M, Robins JM, Hernán MA. Definition and Evaluation of the Monotonicity Condition for Preference-based Instruments. *Epidemiology* 2015.

(3) Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006;17:360-372.

(4) Imbens GW, Angrist JD. Identification and Estimation of Local Average Treatment Effects. *Econometrica* 1994;62:467-475.

(5) Small DS, Tan Z. A stochastic monotonicity assumption for the instrumental variables method. Working Paper, Department of Statistics, University of Pennsylvania, 2007.

(6) Small DS, Tan Z, Lorch SA, Brookhart MA. Instrumental variable estimation when compliance is not deterministic: the stochastic monotonicity assumption. 2014. arXiv: 1407.7308v2

(7) Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *Int J Biostat* 2007;3: Article 14.

(8) den Elzen WP, Lefebre-van de Fliert AA, Virgini V et al. International variation in GP treatment strategies for subclinical hypothyroidism in older adults: a case-based survey. *Br J Gen Pract* 2015;65:e121-e132.

(9) Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 1996;91:444-455.

(10) Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf* 2010;19:537-554.

(11) *PLOTMATRIX: Stata module to plot values of a matrix as different coloured blocks.* [Version S439602. Boston College Department of Economics; 2004.

(12) Swanson SA, Hernán MA. Think globally, act globally: An epidemiologist's perspective on instrumental variable estimation. *Stat Sci* 2014;29:371-374.

(13) Imbens GW. Instrumental Variables: An Econometrician's Perspective. *Stat Sci* 2014;29:323-358.

eTable 1. *Response rates per country and overall*

Country	Responses	Surveys sent out	Response rate (%)
The Netherlands	129	315	41
United Kingdom	22	178	34
New Zealand	31	850	4
Ireland	21	150	14
Switzerland	262	1086	25
Germany	61	178	34
Total	526	2710	19

Adapted from Den Elzen et al, British Journal of General Practice 2015.

eTable 2. *Probability of levothyroxine dependent on treatment of all other cases by the same general practitioner (GP).*

Case	Country															
	The Netherlands (n=117)								Switzerland (n=235)							
	0/7	1/7	2/7	3/7	4/7	5/7	6/7	7/7	0/7	1/7	2/7	3/7	4/7	5/7	6/7	7/7
1	0	0	13	25	4	47	73	67	0	33	7	53	15	48	62	86
2	0	0	0	12	1	42	81	100	0	14	12	22	6	37	73	94
3	5	50	75	92	76	100	100	100	6	67	47	98	96	100	100	100
4	0	44	86	100	86	100	84	98	0	45	39	98	88	88	100	100
5	0	0	0	8	8	46	62	87	0	0	6	27	5	13	48	94
6	0	0	0	21	10	28	46	80	3	17	12	36	6	26	53	89
7	3	33	64	87	69	93	100	100	6	64	18	92	80	93	93	100
8	5	50	70	87	58	75	79	80	3	58	36	95	74	88	100	100

Percentage of yes answers per case within each country, dependent on the treatment of the other cases of the same GP. The column headings indicate the proportion of the other patients treated.

Chapter 3

Physician's preference-based instrumental variable analysis: is it valid and useful in a moderate-sized study?

Anna G.C. Boef*, Judith van Paassen*, M. Sesmu Arbous, Arno Middelkoop, Jan P. Vandenbroucke, Saskia le Cessie, Olaf M. Dekkers
* These authors contributed equally to this work

Epidemiology. 2014 Nov;25(6):923-7

Abstract

Background: Instrumental variable methods can potentially circumvent the unmeasured confounding inherent in observational data analyses.

Methods: We investigated the validity and usefulness of physician's preference instrumental variable analysis in the setting of a moderate-sized clinical study. Using routine care data from 476 elective cardiac surgery patients, we assessed the effect of preoperative corticosteroids on mechanical ventilation time and duration of intensive care and hospital stay, occurrence of infections, atrial fibrillation, heart failure and delirium.

Results: Although results of the physician's preference-based instrumental variable analysis corresponded in direction to results of a recent large randomized trial of the same therapy, the instrumental variable estimates showed much larger effects with very wide confidence intervals.

Conclusion: The lesser statistical precision limits the usefulness of instrumental variable analysis in a study that might be of sufficient size for conventional analyses, even if a strong and plausible instrument is available.

Instrumental variable analysis can potentially circumvent confounding by indication that exists due to unknown or poorly recorded factors in observational data of anticipated therapy effects.[1] Physician's prescribing preference is a promising instrument, since differences among physicians in therapy preferences are ubiquitous. We used anesthesiologist's preference in an instrumental variable analysis to investigate whether preoperative high-dose corticosteroids are beneficial in cardiac surgery patients because they suppress the procedure-induced inflammatory response.[2;3] We compared instrumental variable analyses to standard regression techniques, and also to results from the recent Dexamethasone for Cardiac Surgery randomized trial.[4]

Methods

We used clinical data collected in the context of routine clinical care. The Leiden University Medical Centre review board waived the need of formal ethical approval and written informed consent.

Study population.
We assessed data on all adult patients who underwent elective cardiac surgery in the Leiden University Medical Centre in 2005. Patients had undergone a range of interventions, including coronary artery bypass grafting, valve repair/replacement, and heart failure surgery. Patients treated with corticosteroids prior to admission for cardiac surgery were excluded, leaving 476 patients, of whom 115 received prophylactic corticosteroids. All received regular care according to the fast-track protocol.[5]

Study end points.
Data on demographic features, type of surgical intervention and EuroSCORE were extracted from electronic and paper patient records. The EuroSCORE is a validated prognostic score of in-hospital mortality, based on patient-related, cardiac-related, and operation-related factors.[6;7] Primary endpoints were 30-day mortality, ventilation time, and durations of intensive care unit (ICU) and hospital stays. Secondary outcomes were atrial fibrillation, infections, heart failure, delirium, norepinephrine use, glucose and leukocyte count

Statistical analysis
We first used linear regression to estimate the effect of corticosteroids on the outcomes. This included crude analyses, multivariable analyses (adjusting for age, sex, diabetes, EuroSCORE and type of surgery) and propensity score-adjusted analyses (including the variables in the multivariable model plus the surgeon). Next, we performed two-

stage least squares instrumental variable analysis, with robust standard errors for dichotomous outcomes. The instrument was the proportion of all earlier patients of the same anesthesiologist who received corticosteroids. We selected this instrument based on the first-stage F-statistic and partial r^2 and on the range of predicted treatment probabilities. IV analyses were based on 461 patients (excluding 3 patients with unknown anesthesiologist, the only 2 patients of one anesthesiologist, and all first patients of the 10 anesthesiologists). Instrumental variable assumptions for our study were as follows: (1) anesthesiologist's preference affects the probability that a patient receives corticosteroids; (2) anesthesiologist's preference for corticosteroids does not affect the outcome other than through the decision whether to administer corticosteroids and (3) anesthesiologist's preference for corticosteroids is not related to characteristics of his patient population.[8;9] The fourth assumption, required to obtain a point estimate,[10;11] was the monotonicity assumption: no anesthesiologist would give corticosteroids to a certain patient unless all anesthesiologists with the same or a stronger preference would also give corticosteroids to that patient. The causal effect estimated is a local average treatment effect,[11] a weighted average of the treatment effects in patients who would receive corticosteroids from anesthesiologists with a certain preference level, but not from anesthesiologists with a lower preference.[10] Statistical analyses were performed with Stata 12 and the extension ivreg2.[12]

For additional information regarding study population, data-extraction, study endpoints, conventional analyses, instrumental variable analyses and sensitivity analyses, see the eAppendix.

Results

Table 1 displays patient characteristics and outcomes according to received treatment. The EuroSCORE was higher in patients who received corticosteroids, suggesting confounding.

For the selected instrument the first stage F-statistic was 126 and the partial r^2 was 0.22 (see eMethods and eTable 1). Table 2 shows patient characteristics across physician's preference quintiles. There was no clear pattern across physician's preference quintiles in EuroSCORE (see eFigure 2 for EuroSCORE per anesthesiologist) or other patient characteristics, suggesting physicians' preference for corticosteroids was not related to differences in patients' prognosis. Table 2 shows a decreasing pattern across physician's preference categories for duration of ventilation and infections.

Results of conventional and instrumental variable analyses are displayed in the Figure (dichotomous outcomes only) and eTable 2. In general, unadjusted conventional analyses showed poorer outcomes in patients treated with corticosteroids (except for

Table 1. *Patient characteristics and outcomes[a] by treatment status.*

	Prophylactic corticosteroids			
	No (n=361)		Yes (n=115)	
Patient Characteristics				
Male	246	(68)	69	(60)
Age (years); mean (SD)	64.5	(13.5)	63.9	(12.9)
BMI (kg/m^2); mean (SD)	26.6	(4.2)	26.4	(4.2)
Diabetes mellitus	54	(15)	15	(13)
EuroSCORE; median (IQR)	4	(2-8)	5	(3-10)
EuroSCORE category				
1-2%	115	(32)	23	(20)
3-5%	110	(31)	35	(31)
≥6%	134	(37)	55	(49)
Type of surgery				
Off-pump CABG	36	(10)	6	(5)
On-pump CABG	100	(28)	29	(25)
Valve	116	(32)	39	(34)
Combination/ Other	109	(30)	41	(36)
Outcomes				
Mortality (30 days)	10	(2.8)	4	(3.5)
Ventilation time (hrs); median (IQR)	10	(7-19)	11	(7-20)
ICU stay (days); median (IQR)	1	(1-3)	2	(1-4)
Hospital stay (days); median (IQR)	7	(6-11)	8	(6-13)
Highest norepinephrine dose > 0.1μg/kg/min	112	(33)	35	(32)
Highest glucose (mmol/l); mean (SD)	10.4	(2.4)	11.4	(2.5)
Highest leukocyte count (10^9/L); mean (SD)	13.4	(4.0)	15.6	(5.1)
Atrial fibrillation	173	(48)	50	(44)
Infection	52	(15)	15	(13)
Heart failure	48	(13)	22	(19)
Delirium	54	(15)	20	(18)

Abbreviations: BMI, body mass index; CABG coronary artery bypass graft; IQR, interquartile range

[a] No. (%), unless otherwise indicated

Table 2. *Patient characteristics and outcomes[a] by instrumental variable status.*

	Instrumental variable quintile				
	1 (n=104)	2 (n=84)	3 (n=89)	4 (n=92)	5 (n=92)
% Prophylactic corticosteroids in all previous patients	0	2-8	9-19	19-42	43-100
Prophylactic corticosteroids in current patient	8 (8)	3 (4)	19 (21)	27 (29)	57 (62)
Patient Characteristics					
Male	69 (66)	52 (62)	59 (66)	63 (68)	61 (66)
Age (years); mean (SD)	62.6 (13.6)	65.2 (12.9)	65.1 (13.6)	64.0 (13.4)	63.8 (13.1)
BMI (kg/m^2); mean (SD)	25.9 (3.7)	26.5 (4.2)	26.4 (4.2)	26.3 (3.8)	27.1 (4.6)
Diabetes mellitus	14 (14)	9 (11)	13 (15)	11 (12)	16 (17)
EuroSCORE; median (IQR)	4 (2-9)	4 (2-8)	5 (3-8)	4 (2-9)	4 (2-9)
EuroSCORE category					
1-2%	34 (33)	28 (33)	18 (20)	29 (32)	28 (31)
3-5%	27 (26)	22 (26)	28 (32)	28 (30)	32 (35)
≥6%	41 (40)	34 (40)	42 (48)	35 (38)	31 (34)
Type of surgery					
Off-pump CABG	8 (8)	10 (12)	9 (10)	8 (9)	7 (8)
On-pump CABG	26 (25)	25 (30)	15 (17)	27 (29)	28 (30)
Valve	32 (31)	30 (36)	35 (39)	23 (25)	32 (35)
Combination/ Other	38 (37)	19 (23)	30 (34)	34 (37)	25 (27)
Outcomes					
Mortality (30 days)	2 (1.9)	2 (2.4)	5 (5.6)	4 (4.4)	1 (1.1)
Ventilation time (hrs); median (IQR)	12 (8-20)	11 (7-23)	12 (7-24)	10 (7-20)	9 (6-14)
ICU stay (days); median (IQR)	2 (1-4)	1 (1-3)	2 (1-3)	2 (1-4)	1 (1-1)
Hospital stay (days); median (IQR)	8 (6-13)	7 (5-9)	8 (6-11)	8.5 (6-14)	8 (5.5-9)
Highest norepinephrine dose > 0.1µg/kg/min	28 (31)	30 (37)	32 (39)	39 (45)	14 (16)
Highest glucose (mmol/l); mean (SD)	10.6 (2.3)	10.6 (2.5)	10.6 (3.0)	10.2 (1.8)	11.2 (2.6)
Highest leukocyte count (10^9/L); mean (SD)	13.0 (3.3)	12.8 (3.3)	14.7 (5.2)	14.4 (4.9)	15.0 (4.8)
Atrial fibrillation	42 (41)	40 (48)	45 (51)	44 (48)	44 (48)
Infection	18 (18)	16 (19)	12 (13)	12 (13)	9 (10)
Heart failure	18 (17)	9 (11)	19 (21)	14 (15)	9 (10)
Delirium	16 (16)	15 (18)	16 (18)	16 (18)	9 (10)

[a] No. (%), unless otherwise indicated

atrial fibrillation, infections and norepinephrine dose). Multivariable and propensity-score-adjusted analyses generally showed a null effect. Instrumental variable results indicated a decreased risk of adverse outcomes (except atrial fibrillation) after corticosteroid administration. However, confidence intervals of IV estimates were much wider than those of conventional estimates. For example, crude analysis indicated the risk of a ventilation time >11 hours was 3.1% higher (95% confidence interval = -7.8% to 14.1%), propensity-score-adjusted analysis indicated it was 2.0% lower (-12.8% to 8.8%) and instrumental variable analysis indicated it was 28.1% lower (-52.4% to -3.9%) for patients who received corticosteroids. Instrumental variable estimates of differences in glucose and leukocyte count were slightly higher than estimates from the other analyses (eTable 2).

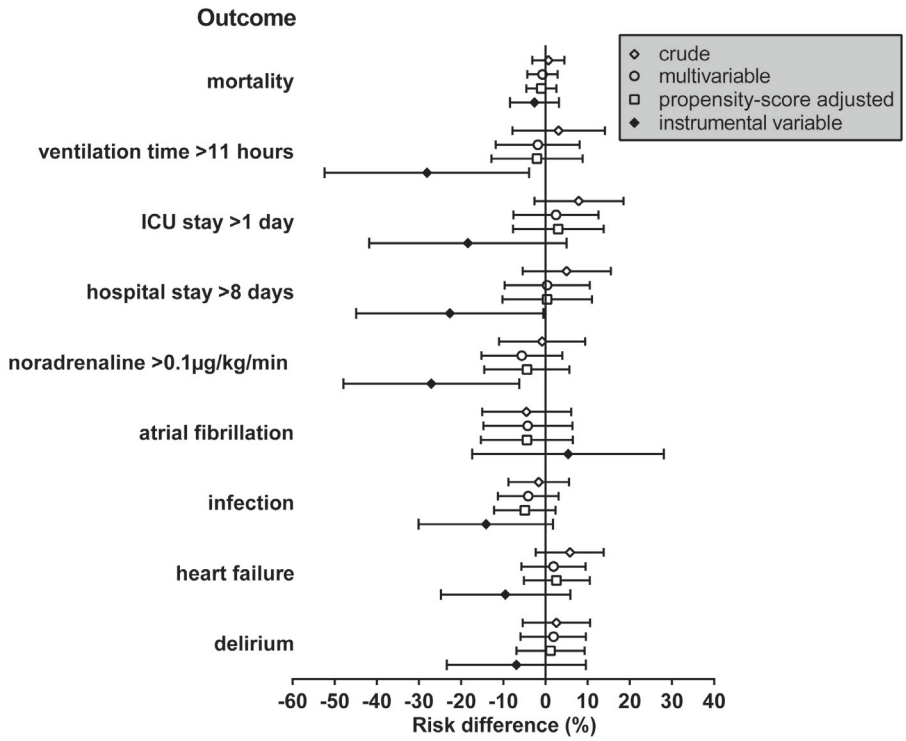


Figure 1 *Estimates of the effect of prophylactic corticosteroids on clinical outcomes in cardiac surgery patients, from crude, multivariable, propensity-score-adjusted and instrumental variable analyses. Risk differences with 95% confidence interval are shown.*

Due to our small sample size we could compare our results only to secondary outcomes of the Dexamethasone for Cardiac Surgery randomized clinical trial.[4] In general, effects in our instrumental variable analyses were similar in direction to the randomized clinical trial results (see eResults), but with considerably larger effect sizes. For example, whereas our instrumental variable analyses estimated the risk of a ventilation time >24 hours to be 16.3% lower (-33.2% to 0.5%) for patients who received corticosteroids, the randomized clinical trial estimated this difference to be -1.5% (-2.7% to -0.3%).[4]

Neither adjusting the instrumental variable analysis for patient characteristics, nor using an instrumental variable based on the last 5 patients materially changed the results (eTable 3). Sensitivity analyses estimating relative risks yielded similar effect sizes (eTable 4).

Discussion

We investigated whether physician's preference-based instrumental variable analysis was valid and useful in a moderate-sized study for the question whether preoperative corticosteroids are beneficial in cardiac surgery. In contrast to crude and propensity score adjusted analyses, instrumental variable analysis using anesthesiologists' preferences as an instrument showed beneficial effects, similar in direction to the Dexamethasone for Cardiac Surgery randomized clinical trial results,[4] and compatible with pathophysiologic insights concerning prevention of operation-induced systemic inflammation.[13-15] However, compared with the trial results, the instrumental variable estimates were extremely large and confidence intervals were so wide as to preclude useful conclusions.

A reason for the difference in magnitude between our instrumental variable estimates and the randomized clinical trial results could be effect modification due to baseline prognostic differences between the study populations. Our patients seemed to be more high risk, as indicated by longer ventilation and ICU stay times and higher incidences of most outcomes.

There are also design-inherent explanations for the large size of the instrumental variable effect estimates. First, our smaller number of patients, compared with the randomized clinical trial, gives rise to less statistical precision, which is further aggravated in the IV analysis due to its two-stage approach.[16] This lack of precision, reflected in the large confidence intervals, could lead to the instrumental variable estimates being more extreme by chance.

Second, main instrumental variable assumptions may be violated. We would not expect differences in patient characteristics depending on anesthesiologist's

preference for corticosteroids (independence assumption), as patients are assigned to the anesthesiologist on duty on the day of surgery. The lack of a consistent pattern in measured patient characteristics across quintiles of the instrumental variable is therefore reassuring. The assumption that preference for corticosteroids does not affect outcomes other than through administration of corticosteroids is more difficult to assess but seems plausible, as anesthesiologists took care of the patients only during surgery and were not involved in subsequent ICU care.

Third, violation of the monotonicity assumption could contribute to the extreme estimates. For example, if patients who receive corticosteroids from an anesthesiologist with a weak preference would not receive them from an anesthesiologist with a strong preference and if corticosteroids are of relatively little benefit to these patients, then the estimate of the effect of corticosteroids would be too favorable.

Fourth, estimands of the conventional and the instrumental variable analyses are different: the conventional analyses estimate average treatment effects in the population, while the instrumental variable analyses estimate local average treatment effects (as explained in the Methods section).

Fifth, finite sample bias might be a reason for the large instrumental variable effect estimates. However, the first-stage F-statistic of 126 should be sufficient for finite sample bias to be negligible.[1] We further explored this using simulations under conditions similar to our study (100 to 500 patients; mean partial r^2 of 0.17; unmeasured confounding and a binary outcome occurring in 50% of patients) (see eAppendix). Mean instrumental variable estimates were close to the "true" treatment effect of -0.10, even when the sample size was reduced to 100 patients, indicating no substantial finite sample bias with an instrument of this strength.

In conclusion, despite availability of a strong instrument, plausibly fulfilling main instrumental variable assumptions, physician's preference-based instrumental variable analysis in a moderate-sized study population showed results that differed greatly in magnitude from results of a major randomized clinical trial on the same intervention. We have explored possible reasons and conclude that this phenomenon is most likely due to the reduced statistical precision of the instrumental variable analysis in datasets of moderate size.

References

(1) Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. *Epidemiology* 2006;17:260-267.

(2) Ng CS, Wan S, Arifi AA, Yim AP. Inflammatory response to pulmonary ischemia-reperfusion injury. *Surg Today* 2006;36:205-214.

(3) Clark SC. Lung injury after cardiopulmonary bypass. *Perfusion* 2006;21:225-228.

(4) Dieleman JM, Nierich AP, Rosseel PM et al. Intraoperative high-dose dexamethasone for cardiac surgery: a randomized controlled trial. *JAMA* 2012;308:1761-1767.

(5) Silbert BS, Santamaria JD, O'Brien JL, Blyth CM, Kelly WJ, Molnar RR. Early extubation following coronary artery bypass surgery: a prospective randomized controlled trial. The Fast Track Cardiac Care Team. *Chest* 1998;113:1481-1488.

(6) Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg* 1999;16:9-13.

(7) Roques F, Michel P, Goldstone AR, Nashef SA. The logistic EuroSCORE. *Eur Heart J* 2003;24:881-882.

(8) Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* 2006;17:268-275.

(9) Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. *J Clin Epidemiol* 2009;62:1226-1232.

(10) Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006;17:360-372.

(11) Swanson SA, Hernan MA. Commentary: how to report instrumental variable analyses (suggestions welcome). *Epidemiology* 2013;24:370-374.

(12) Baum CF, Schaffer ME, Stillman S. ivreg2: Stata module for extended instrumental variables/2SLS, GMM and AC/HAC, LIML and k-class regression. 2010.

(13) Warren OJ, Smith AJ, Alexiou C et al. The inflammatory response to cardiopulmonary bypass: part 1--mechanisms of pathogenesis. *J Cardiothorac Vasc Anesth* 2009;23:223-231.

(14) Aird WC. The role of the endothelium in severe sepsis and multiple organ dysfunction syndrome. *Blood* 2003;101:3765-3777.

(15) Marshall JC. Inflammation, coagulopathy, and the pathogenesis of multiple organ dysfunction syndrome. *Crit Care Med* 2001;29:S99-106.

(16) Ionescu-Ittu R, Delaney JA, Abrahamowicz M. Bias-variance trade-off in pharmacoepidemiological studies using physician-preference-based instrumental variables: a simulation study. *Pharmacoepidemiol Drug Saf* 2009;18:562-571.

eMethods

1. Study population (additional information)

During the study period prophylactic corticosteroids were not routinely administered to cardiac surgery patients in the LUMC. Data on administration of prophylactic corticosteroids were extracted from the automated registration system for the operating room in which all administered medications were registered. Of the 476 patients in this study, 115 received prophylactic corticosteroids. Of the 361 patients who did not receive prophylactic corticosteroids, 73 did receive corticosteroids during the surgical procedure or as treatment of protamine-allergy at the end of surgery. After surgical intervention all patients were admitted to the cardio-thoracic intensive care unit (ICU).

2. Data extraction and study end-points

Data were extracted from electronic patient record databases, routinely used in the operating room and in the ICU (Metavision®, Mirador®) in which clinical parameters are collected automatically. In case data was missing in these electronic records, data was extracted from paper patient charts. These were kept simultaneously during the conversion phase from paper patient charts to electronic patient records. We extracted data on demographic features and type of surgical intervention. Furthermore, the logistic EuroSCORE, routinely computed and registered by the thoracic surgery department, was obtained for all patients. This a validated prognostic score of in-hospital mortality related to cardiac surgery, based on patient-related factors (age, sex, chronic pulmonary disease, extra cardiac arteriopathy, neurological dysfunction, previous cardiac surgery, serum creatinine, active endocarditis, critical pre-operative state), cardiac-related factors (unstable angina, left ventricular dysfunction, recent myocardial infarct, pulmonary hypertension) and operation-related factors (emergency, other than isolated CABG, surgery on thoracic aorta, post infarct septal rupture).[1;2] The following clinical study end-points were recorded: 30-day mortality, ventilation time, duration of ICU and hospital stay. These study end-points are collected and checked systematically on a weekly basis by a quality manager and by the hospital billing department. The following clinical parameters were extracted from the electronic patient records and recorded for the study: highest necessary dose of norepinephrine, highest glucose value and highest leukocyte count in the first 24 hours after intervention. The occurrence of atrial fibrillation, infections, heart failure or delirium during hospital stay was also extracted. Infection was defined as clinical symptoms requiring new antibiotic treatment; heart failure was defined as a clinical diagnosis requiring additive diuretic, or invasive supportive (intra-aortic balloon pump, assist device) treatment; delirium was defined as the need for haloperidol.

3. Main assumptions for instrumental variable analyses

In order to be valid, an instrumental variable should fulfill three main assumptions, which we will discuss specifically applied to our study (see also eFigure 1). The first assumption is that anesthesiologist's preference affects the probability that a patient receives corticosteroids. The second assumption is that the anesthesiologist's preference for corticosteroids does not affect the outcome in other ways than through the decision of whether to administer corticosteroids (exclusion restriction); the third is that the anesthesiologist's preference for corticosteroids is not related to characteristics of his patient population (independence assumption).[3;4] The difference in outcomes can then be attributed entirely to the difference in the probability of receiving corticosteroids (based on the anesthesiologist's preference).

We explored whether there was variation in corticosteroid administration amongst anesthesiologists and whether this seemed independent of their patient population. The lower part of eFigure 2 shows the proportion of patients to whom the anesthesiologists administered prophylactic corticosteroids; the upper part shows box plots of the EuroSCORE of these patients. The percentage of patients to whom the anesthesiologists administered corticosteroids showed considerable variation, ranging from 0% to 63%. In our data, there is no consistent pattern in the EuroSCORE with increasing prescription of corticosteroids (in accordance with the independence assumption), giving general reassurance that we could use anesthesiologist's preference as an instrumental variable

4. Instrumental variable selection

In our study population there was large variation among anesthesiologists in frequency of administration of prophylactic corticosteroids, ranging from 0% to 63%. This indicated that anesthesiologist's preference regarding administration of prophylactic corticosteroids was a potentially suitable instrument. We considered several estimates of anesthesiologist's preference for use as an instrument, based on one, two, five, ten or all previous patients. For a given patient the proportions of these preceding patients who received prophylactic corticosteroids were calculated, to provide estimates of the anesthesiologist's relative preference for prophylactic corticosteroids at the time of the treatment decision for this specific patient.

To identify which of our candidate instruments was most strongly related to treatment, we carried out the first stage of the two-stage least squares instrumental variable regression only, by means of linear regression of the treatment on the candidate instrument.[5] We selected the strongest instrumental variable based on the F-statistic and partial r^2 of the first stage of the two-stage least squares regression and on the range

of predicted probabilities of treatment. An F-statistic greater than 10 suggests that small sample bias is negligible and that the instrument is therefore sufficiently strong.[6] The partial r^2 indicates which proportion of the variance of the treatment is explained by the instrumental variable.[7]

eTable 1 displays the regression coefficients, the F-statistic and the partial r^2 for the first stage regression using each of the candidate instruments. The regression coefficient can be interpreted as follows for the instrument based on the last patient only: for a patient treated by an anesthesiologist who administered corticosteroids to the previous patient the probability of receiving corticosteroids was 0.28 higher than for a patient treated by an anesthesiologist who did not administer corticosteroids to the previous patient. Analogously, for a patient treated by an anesthesiologist who administered corticosteroids to all previous patients the expected probability of receiving corticosteroids would be 0.82 higher than for a patient treated by an anesthesiologist who administered corticosteroids to none of their previous patients. The strengths of instrumental variables based on 10 previous prescriptions or all previous prescriptions were very similar, with a partial r^2 of 0.21 and 0.22 and F-statistics of 131 and 126 respectively. These instruments were considerably stronger than the instruments based on just one or two previous prescriptions. Although the partial r^2 and F-statistic were slightly higher for the instrument based on 10 previous prescriptions than for the instrument based on all previous prescriptions, the range of predicted probabilities of treatment was slightly larger for the latter instrument. We therefore selected the proportion of all previous patients who received prophylactic corticosteroids for use as an instrument in subsequent analyses.

5. Conventional statistical analyses

Crude analysis

For continuous outcomes we calculated a mean difference (MD) with 95% confidence interval (CI) between treatment groups. For binary outcomes we calculated a risk difference (RD) with 95% CI, because this effect measure can be compared directly to two-stage least squares instrumental variable results. Ventilation time in hours and duration of ICU and hospital stay in days were dichotomized (as shorter or longer than the median). Robust standard errors were used for dichotomous outcomes.

Multivariable model and propensity score adjusted analyses

The above analyses were repeated using multivariable adjustment and propensity score adjustment. The multivariable model was adjusted for age, sex, diabetes mellitus, EuroSCORE and type of surgical procedure, for the 470 patients with information on all included covariates. Operating surgeon was not included in the multivariate

regression models because for many outcomes there were few events, limiting the number of covariates that can be included in the regression model. The propensity score was calculated by first performing a logistic regression model with receipt of prophylactic corticosteroids as the dependent variable and all variables used in the multivariable model plus the operating surgeon as covariates and then predicting the probabilities of treatment for each patient based on this model. This was done for the 464 patients with information on all variables.

6. Sensitivity analyses

(1) an instrumental variable analysis adjusted for age, sex, EuroSCORE, type of intervention and diabetes, to explore the effect of additional adjustments.

(2) an instrumental variable analysis using an alternative instrument based on treatment of the previous 5 patients only, which might accommodates preference changes over time better than an instrument based on all previous patients.

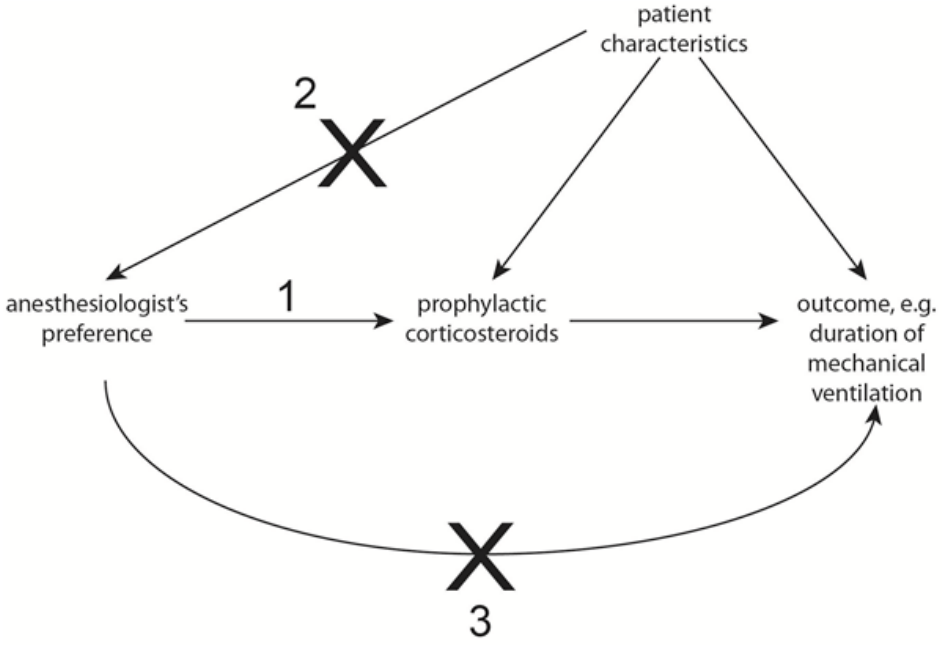
(3) an analysis in which we replaced the second stage of the instrumental variable regression with a generalized linear model with a log-link, which gives relative risk estimates, because two-stage least squares regression is based on linear models and may pose problems if exposures and outcomes are dichotomous, including predicted values below 0 or above 1.

7. Timeline of analyses relative to the DECS randomized trial[8]

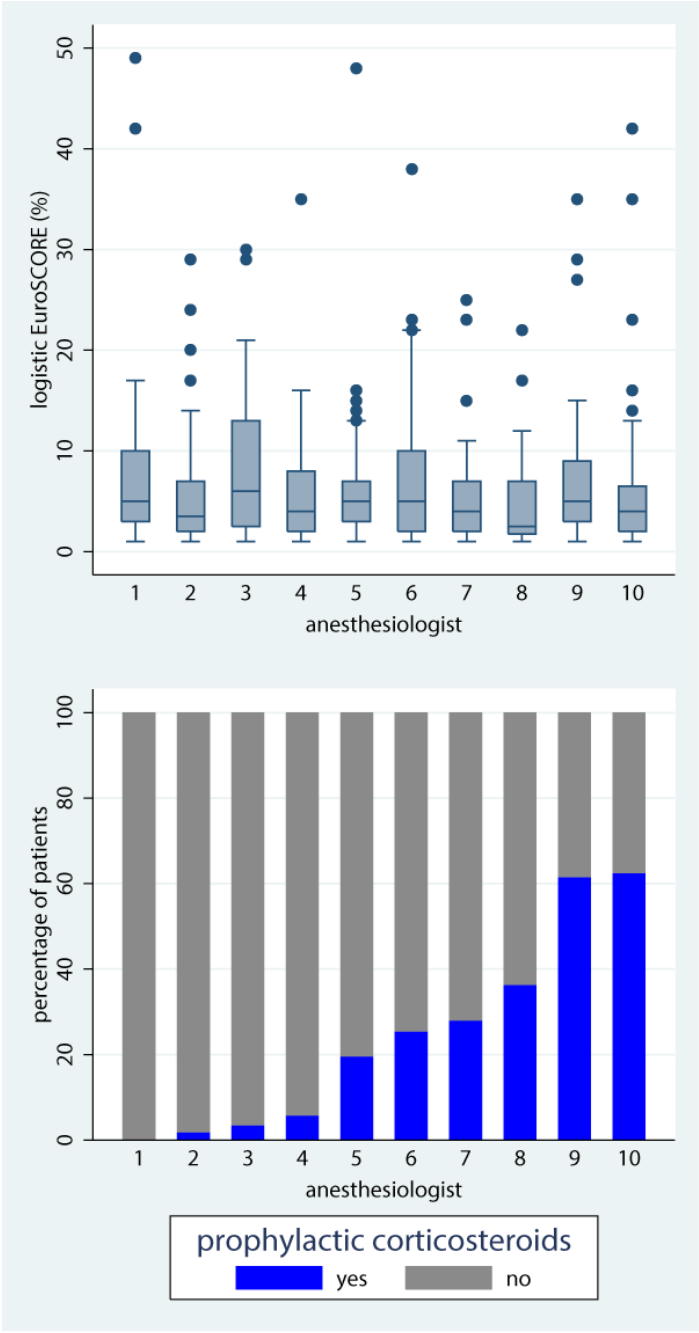
All major decisions about patient selection, choice of instrument, outcomes and types of analysis were made before we knew the DECS trial results, to which we compared our results. After the trial was published we performed additional analyses with cut-off points similar to those in the trial for a better comparison.

References

(1) Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg* 1999;16:9-13.
(2) Roques F, Michel P, Goldstone AR, Nashef SA. The logistic EuroSCORE. *Eur Heart J* 2003;24:881-882.
(3) Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* 2006;17:268-275.
(4) Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. *J Clin Epidemiol* 2009;62:1226-1232.
(5) Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf* 2010;19:537-554.
(6) Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. *Epidemiology* 2006;17:260-267.
(7) Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables II: instrumental variable application-in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. *J Clin Epidemiol* 2009;62:1233-1241.
(8) Dieleman JM, Nierich AP, Rosseel PM et al. Intraoperative high-dose dexamethasone for cardiac surgery: a randomized controlled trial. *JAMA* 2012;308:1761-1767.



eFigure 1. *Instrumental variable assumptions in this study.*
Causal diagram depicting instrumental variable assumptions if anesthesiologist's preference is used as an instrumental variable in a study investigating the effect of prophylactic corticosteroids in cardiac surgery patients.



eFigure 2.
Proportion of patients to whom an anesthesiologist administered prophylactic corticosteroids (lower part) and distribution of the EuroSCORE of these patients (upper part).

eTable 1. *Strength of instruments based on 5 different preference assignments.*

Instrument	Difference in probability of treatment (95% CI)*	*F*-statistic	Partial r^2
Previous patient†	0.28 (0.19-0.37)	39	0.08
Last 2 patients‡	0.47 (0.36-0.58)	74	0.14
Last 5 patients‡	0.68 (0.56-0.80)	117	0.20
Last 10 patients‡	0.77 (0.64-0.91)	131	0.22
All previous patients‡	0.82 (0.67-0.96)	126	0.22

†Administration of prophylactic corticosteroids in the previous patient of the same anesthesiologist.

‡Proportion of the last 2/ last 5/ last 10/ all previous patients of the same anesthesiologist who received prophylactic corticosteroids.

*All differences stated are regression coefficients and represent the difference in the probability of receiving prophylactic corticosteroid between patients with values of the instrument of 1 and 0. For the instrument based on the previous patient 1 indicates that the previous patient received the treatment and 0 denotes that the previous patient did not receive the treatment. For the instrument based on all previous patient 1 would denote that all previous patients received the treatment, and 0 would denote that none of the previous patients received the treatment.

eTable 2. *Outcomes by treatment status and estimates of the treatment effect from four different analyses.*

	No prophylactic corticosteroids (n=361)	Prophylactic corticosteroids (n=115)	Crude	Multivariable model	Propensity score adjusted	Instrumental variable
Primary outcome†						
Mortality (30 days)	10 (2.8)	4 (3.5)	0.7 (-3.1,4.5)	-0.7 (-4.3,2.9)	-1.0 (-4.5,2.6)	-2.6 (-8.4,3.2)
Ventilation time >11 hrs	158 (46)	52 (49)	3.1 (-7.8,14.1)	-1.8 (-11.8,8.1)	-2.0 (-12.8,8.8)	-28.1 (-52.4,-3.9)
ICU stay >1 day	169 (47)	63 (55)	7.9 (-2.6,18.5)	2.5 (-7.6,12.6)	3.0 (-7.7,13.8)	-18.4 (-41.8,5.0)
Hospital stay >8 days	139 (39)	50 (44)	5.0 (-5.4,15.5)	0.4 (-9.7,10.5)	0.4 (-10.2,11.0)	-22.7 (-44.9,-0.5)
Ventilation time >24 hrs*	60 (17)	21 (20)	2.4 (-6.3,11.0)	-1.1 (-8.9,6.6)	-1.9 (-9.8,6.0)	-16.3 (-33.2,0.5)
ICU stay>2 days*	115 (32)	43 (38)	5.5 (-4.7,15.7)	1.9 (-8.1,11.9)	1.7 (-8.7,12.0)	-16.2 (-37.2,4.8)
Clinical parameters						
Highest norepinephrine dose > 0.1μg/kg/min †	112 (33)	35 (32)	-0.8 (-11.0,9.4)	-5.6 (-15.2,4.0)	-4.4 (-14.5,5.7)	-27.1 (-47.9,-6.2)
Highest glucose (mmol/l) ‡	10.4 (0.14)	11.4 (0.25)	0.96 (0.40,1.52)	0.82 (0.27,1.37)	0.83 (0.27,1.40)	0.94 (-0.20,2.08)
Highest leukocyte count (10^9/L) ‡	13.4 (0.23)	15.6 (0.51)	2.29 (1.30,3.27)	2.30 (1.29,3.30)	2.33 (1.32,3.35)	3.01 (1.01,5.00)
Complications †						
Atrial fibrillation	173 (48)	50 (44)	-4.5 (-15.0,6.1)	-4.2 (-14.7,6.4)	-4.4 (-15.3,6.5)	5.4 (-17.4,28.1)
Infection	52 (15)	15 (13)	-1.6 (-8.8,5.6)	-4.1 (-11.3,3.1)	-4.9 (-12.2,2.4)	-14.1 (-30.1,1.8)
Heart failure	48 (13)	22 (19)	5.8 (-2.3,13.8)	1.9 (-5.7,9.5)	2.6 (-5.1,10.5)	-9.5 (-24.8,5.9)
Delirium	54 (15)	20 (18)	2.6 (-5.4,10.6)	1.9 (-5.9,9.6)	1.2 (-6.9,9.3)	-6.9 (-23.4,9.6)

† n (%), risk difference in % (95% CI)
‡ mean (SE), mean difference (95% CI)
*Additional analyses for comparison to RCT results.

eTable 3. *Sensitivity analyses.*

	Instrumental variable: unadjusted	Instrumental variable: adjusted[a]	Instrumental variable (last 5)[b]
Primary outcome†			
Mortality (30 days)	-2.6 (-8.4,3.2)	-1.9 (-7.7,4.0)	-5.1 (-10.5,0.3)
Ventilation time >11 hrs	-28.1 (-52.4,-3.9)	-26.1 (-48.3,-4.0)	-28.7 (-54.1,-3.4)
ICU stay >1 day	-18.4 (-41.8,5.0)	-16.6 (-38.4,5.1)	-17.2 (-41.0,6.6)
Hospital stay >8 days	-22.7 (-44.9,-0.5)	-22.7 (-44.6,-0.8)	-27.5 (-50.3,-4.6)
Ventilation time >24 hrs	-16.3 (-33.2,0.5)	-14.7 (-30.3,1.0)	-18.7 (-35.5,-2.0)
ICU stay>2 days	-16.2 (-37.2,4.8)	-13.2 (-33.1,6.7)	-24.0 (-45.7,-2.4)
Clinical parameters			
Highest norepinephrine dose† > 0.1µg/kg/min	-27.1 (-47.9,-6.2)	-26.5 (-45.9,-7.2)	-28.1 (-49.9,-6.4)
Highest glucose (mmol/l)‡	0.94 (-0.20,2.08)	0.82 (-0.28,1.92)	0.40 (-0.78,1.57)
Highest leukocyte count (10^9/L)‡	3.01 (1.01,5.00)	3.15 (1.15,5.15)	3.09 (1.01, 5.16)
Complications †			
Atrial fibrillation	5.4 (-17.4,28.1)	8.4 (-13.9,30.7)	11.7 (-12.1,35.5)
Infection	-14.1 (-30.1,1.8)	-16.1 (-32.3,0.0)	-10.4 (-26.5,5.7)
Heart failure	-9.5 (-24.8,5.9)	-8.6 (-23.1,5.8)	-17.5 (-32.8,-2.2)
Delirium	-6.9 (-23.4,9.6)	-8.2 (-24.6,8.2)	-6.2 (-23.3,10.8)

† risk difference in % (95% CI)

‡ mean difference (95% CI)

a. Instrumental variable analysis adjusted for age, sex, EuroSCORE, type of intervention and diabetes.

b. Instrumental variable analysis using the proportion of the last 5 patients treated with corticosteroids as an instrument.

eTable 4. *Relative risk estimates.*

	RR (95% CI*)
Primary outcome	
Mortality (30 days)	0.38 (0.03,3.16)
Ventilation time >11 hrs	0.50 (0.25,0.89)
ICU stay >1 day	0.68 (0.42,1.08)
Hospital stay >8 days	0.56 (0.26,0.93)
Ventilation time >24 hrs	0.36 (0.10,1.03)
ICU stay>2 days	0.61 (0.27,1.16)
Clinical parameters	
Highest norepinephrine dose > 0.1μg/kg/min	0.42 (0.18,0.80)
Complications	
Atrial fibrillation	1.12 (0.69,1.71)
Infection	0.33 (0.06,1.21)
Heart failure	0.50 (0.12,1.31)
Delirium	0.62 (0.14,1.84)

Relative risk estimates obtained using a two-stage model with a linear first stage and a generalised linear model with log-link second stage. The instrumental variable used was the proportion of all previous patients treated with corticosteroids. Confidence intervals were obtained using a bootstrap procedure with 1000 samples, bias corrected.

eResults. Comparison to DECS trial results.

The RCT found 3.4% of patients in the dexamethasone group and 4.9% in the placebo group had a ventilation time >24 hours, a difference of -1.5% (95% CI -2.7%,-0.3%). The percentage of patients with an ICU stay >48 hours was 10.2% in the dexamethasone group and 14.0% in the placebo group, a difference of -3.8% (95%CI -5.7%,-1.9%). For atrial fibrillation the percentages were 33.1% and 35.2% respectively, a difference of -2.1% (95% CI -4.9%, 0.7%); for infections 9.8% and 14.8%, a difference of -5.3% (95% CI -7.2%, -3.4%); for delirium 9.2% and 11.7%, a difference of -2.5% (95% CI -4.3%, -0.7%).[1] In general the effects on these outcomes were similar in direction to the results of our instrumental variable analyses, but with considerably smaller effect sizes.

References

(1) Dieleman JM, Nierich AP, Rosseel PM et al. Intraoperative high-dose dexamethasone for cardiac surgery: a randomized controlled trial. *JAMA* 2012;308:1761-1767.

eAppendix. Simulation study to investigate the influence of finite sample bias.
Monte Carlo simulations for a series of study population sizes of 100, 200, 300, 400 and 500. The instrument *P* was generated from the standard uniform distribution U(0,1). An unmeasured confounder C_u was generated from the uniform distribution U(0,1).

Treatment *X* was generated from a binomial distribution with individual patients' probabilities of treatment dependent on *P* and C_u according to the following equation:
$P(X=1|P,C_u)= 0.7P + 0.2C_u$

Binary outcome *Y* was generated from a binomial distribution with individual patients' probabilities of the outcome dependent on treatment *X* and on C_u as follows:
$P(Y=1| X,C_u)=0.2 - 0.1X + 0.7C_u$.

Next, the treatment effect was estimated in each sample using ordinary least squares regression and two-stage least squares regression. The mean estimates and their standard deviation for each sample size across 2000 simulations are displayed in the table below. The mean partial r^2 in the simulations was 0.17, slightly lower than in our study data. Even at sample size 100 the mean 2-SLS is very close to the true effect of -0.10, indicating small sample bias is not a concern. However, the 2-SLS estimates are very variable, as indicated by their large standard deviations.

Sample size	OLS estimates, mean (SD)	2-SLS estimates, mean (SD)
100	-0.054 (0.100)	-0.101 (0.270)
200	-0.051 (0.069)	-0.107 (0.182)
300	-0.053 (0.056)	-0.107 (0.145)
400	-0.052 (0.051)	-0.100 (0.128)
500	-0.054 (0.045)	-0.101 (0.111)

Stata code for simulations

```
*create a file in which to store results
drop _all
clear all
postfile simres ssize b1 b2 pr2 F using "filename", replace

*programme for creating one dataset (called "finite")
drop _all
capture program drop finite
```

```
program finite, rclass
drop _all
// ssize = sample size, as a macro
args ssize
//generate patients
set obs `ssize'
gen n=_n
//generate the instrument P
gen P=runiform()
//generation of an unmeasured confounder U
gen U=runiform()
//generation of treatment X
gen PrX= 0.7*P+0.2*U
gen X1 = runiform()
gen X=recode(X1,PrX,1)
recode X (1=0) (else=1)
drop X1
//generation of outcome Y
gen PrY= 0.2-0.1*X+0.7*U
gen Y1 = runiform()
gen Y=recode(Y1,PrY,1)
recode Y (1=0) (else=1)
drop Y1
//ordinary least squares regression
quietly regress Y X
scalar b1 = _b[X]
//two-stage least squares regression
quietly ivreg2 Y (X=P), first
scalar b2 = _b[X]
*also save first stage partial r2 and F-statistic
matrix tmp2 = e(first)
scalar pr2 = tmp2[2,1]
scalar F = tmp2[3,1]
post simres (`ssize') (b1) (b2) (pr2) (F)
end


*run the simulations
```

```
foreach ssize in 100 200 300 400 500{
simulate, reps(2000) seed(312): finite `ssize'
                                }

postclose simres

*analyse the results
use "filename", clear
sort ssize
//calculation of mean and standard deviation of the OLS and 2-SLS estimates
//per sample size across 2000 simulations
by ssize: summarize b1 b2 pr2, detail
```

Chapter 4

Sample size importantly limits the usefulness of instrumental variable methods in epidemiological studies.

Anna G.C. Boef, Olaf M. Dekkers, Jan P. Vandenbroucke and Saskia le Cessie

J Clin Epidemiol. 2014 Nov;67(11):1258-64

Abstract

Objective Instrumental variable (IV) analysis is promising for estimation of therapeutic effects from observational data as it can circumvent unmeasured confounding. However, even if IV assumptions hold, IV analyses will not necessarily provide an estimate closer to the true effect than conventional analyses as this depends on the estimates' bias and variance. We investigated how estimates from standard regression (ordinary least squares (OLS)) and IV (two-stage least squares) regression compare on mean squared error (MSE).

Study Design We derived an equation for approximation of the 'threshold' sample size above which IV estimates have a smaller MSE than OLS estimates. Next, we performed simulations, varying sample size, instrument strength and level of unmeasured confounding. IV assumptions were fulfilled by design.

Results Although biased, OLS estimates were closer on average to the true effect than IV estimates at small sample sizes due to their smaller variance. The 'threshold' sample size above which IV analysis outperforms OLS regression depends on instrument strength and strength of unmeasured confounding, but will usually be large given the typical moderate instrument strength in medical research.

Conclusion IV methods are of most value in large studies if considerable unmeasured confounding is likely and a strong and plausible instrument is available.

Introduction

Conventional methods to estimate therapeutic effects from observational data are often inherently affected by residual confounding due to unmeasured patient risk factors for which they cannot adjust. A potentially promising tool for estimation of therapeutic effects from observational data which may circumvent this problem is instrumental variable (IV) analysis. This method requires the identification of a variable that determines the probability of treatment but is not in other ways associated with the outcome under study and thereby mimics randomization. Expressed more formally, an instrument must fulfil three main assumptions: (1) the instrument is associated with the exposure (treatment); (2) the instrument does not affect the outcome in any other way other than through the exposure (exclusion restriction); (3) the instrument and outcome do not share causes (independence assumption) [1-4]. The above assumptions allow estimation of bounds of the treatment effect [3;5]. One additional assumption which allows a point estimate to be obtained, is the assumption of no heterogeneity of treatment effects, in which case the IV analysis estimates the average treatment effect in the population [3;5]. Note that in case of heterogeneity alternative assumptions can be made, but this is beyond the scope of this paper. Examples of instruments used in studies of therapeutic effects include regional variation in treatment rates (i.e. probability of treatment depends on area of residence) [6] and physician prescribing preference [7-9]. In etiologic studies Mendelian randomisation, which uses genetic information as an IV, is increasingly used [10].

Violations of the exclusion restriction and independence assumption will lead to biased IV estimates [5;7;11]. If those assumptions hold, the IV estimator will be

asymptotically unbiased [1;11]. In contrast, the bias of ordinary least squares (OLS) linear regression depends on the amount of residual confounding. However, whether IV analysis effect estimates can be expected to be closer to the true effect than estimates from conventional analysis depends not only on the bias, but also on the variance of the estimates (larger variances leading to higher probability of deviating estimates). The variance of estimates from IV methods like two-stage least squares (2-SLS) regression is much larger than from linear regression at a given sample size, because IV methods involve two estimation stages instead of one [12].

IV methods have been applied in large pharmaco-epidemiological databases, typically exceeding 10.000 patients. However, study populations in clinical research practice are often much smaller. Although in principle the large variance of the IV estimate at smaller sample sizes may not influence the validity of the IV estimates, it does affect how informative and useful the IV estimate is. It translates into a very wide confidence interval and the mean squared error of the IV estimate may be much larger than that of the biased conventional estimate [1]. The influence of sample size on the error of IV estimates has been investigated in conjunction with violations of IV assumptions [13]. In contrast, we will focus on the ideal scenario in which the exclusion restriction and independence assumptions hold to focus on the role of sample size, confounding and strength of instrument. Using theoretical derivations and simulations we will investigate the influence of sample size on how OLS linear regression estimates and 2-SLS IV regression estimates compare in terms of mean squared error (which incorporates both the bias and the variance of the estimates), depending on instrument strength and level of confounding.

Two-stage least squares instrumental variable analysis

Two-stage least squares (2-SLS) IV regression involves two linear regression steps. The first stage linear regression is used to obtain predicted probabilities of treatment for each patient, based on the instrument. Covariates can be included, giving predicted probabilities of treatment conditional on the instrument and these observed covariates. The independence assumption then states that the instrument is not related to patient prognosis given these covariates [2]. The second stage is a regression of the outcome on these predicted treatment probabilities (and covariates if included), thereby providing an estimate of the effect of the treatment on the outcome [2;7;14]. For continuous outcomes the obtained effect estimate is a mean difference and for binary outcomes a risk difference.

The variance of the 2-SLS estimate is

$$\mathrm{var}(\hat{\beta}_{IV}) = \frac{\sigma^2_{Y.X,C}}{n\sigma^2_{X.C} \cdot \rho^2_{X,Z.C}},$$

where $\rho_{X,Z.C}$ is the partial correlation between the instrument Z and the exposure X given covariates C, i.e. the strength of the instrument, and Y is the outcome [11;15]. The variance is therefore $1/\rho^2_{X,Z.C}$ times larger than the variance of an ordinary least squares linear regression (OLS) estimate. This implies that the confidence interval (CI) for the 2-SLS estimator is $1/\rho_{X,Z.C}$ times wider than the CI of the OLS estimator. For example, for a moderately strong instrument with a correlation between instrument and exposure of 0.2, the CI for the 2-SLS estimator will be 5-fold wider than the CI of the OLS estimator. If IV assumptions hold the 2-SLS estimates are asymptotically unbiased: bias will exist in finite samples and depends on the sample size and strength of the instrument. This is known as small sample bias [5;11], finite sample bias [1] or weak instrument bias [16]. The partial F-statistic of

the first stage regression provides an indication of the magnitude of the small sample bias: generally small sample bias is negligible at an F-statistic above 10 [11].

Mean squared error: a summary measure for bias and variance

The mean squared error (MSE) measures the squared average deviation of an estimated effect from the true effect. It is equal to

$$MSE = E[(\hat{\beta} - \beta)^2] \quad (1)$$

in which E denotes expectation, $\hat{\beta}$ is the estimated treatment effect and β is the true treatment effect. It can be shown that the MSE is the sum of the variance and the squared bias of an estimate. It is a measure of how far on average the effect estimate is from the true effect. Comparison of the MSEs of the different analysis methods therefore indicates which estimate is closest on average to the true effect.

Calculation of a sample size at which IV outperforms OLS on mean squared error

The trade-off between the larger bias of the OLS estimates and the larger variance of the IV estimates means that OLS estimates will be closer on average to the true effect at small sample sizes, but IV estimates will eventually be closer on average to the true effect as sample size increases. We derived Equation 2 (the derivation is provided in eAppendix 1) to calculate the 'threshold' sample size ($n_{threshold}$) above which IV analysis will outperform OLS in terms of mean squared error. The equation is an approximation: it does not include small sample bias of the IV estimates and assumes that IV assumptions are fulfilled.

The threshold sample size equals

$$n_{\text{threshold}} = \frac{\sigma^2_{Y.X,C}}{\sigma^2_{X.C}\,\text{bias}^2_{OLS}}\left(\frac{1}{\rho_{X,Z.C}{}^2} - 1\right), \qquad (2)$$

in which bias_{OLS} is the bias of the OLS estimate. As can be seen from the equation, the more biased the OLS estimator and the stronger the instrument (the higher $\rho_{X,Z.C}$), the lower the sample size at which the IV estimator will outperform the OLS. The observed correlation between the instrument and exposure (treatment) [17] can be used to estimate $\rho_{X,Z.C}$; if adjustments for observed covariates are made, the partial correlation controlled for covariates is used. The partial r^2 (the squared partial correlation) for the first stage of the IV analysis reported in several previous IV studies ranged from 0.004 for a very weak IV to 0.22 for a strong IV (hospital preference)[18-20].

The expression $\text{bias}_{OLS} / \sqrt{\frac{\sigma^2_{Y.X,C}}{\sigma^2_{X.C}}}$ can be viewed as "standardised bias" in units of residual standard deviation of Y per standard deviation of X given covariates C. It is equal to the remaining correlation between the exposure and the outcome due to residual confounding. Especially in case of binary exposures and outcomes the standardised bias will typically be much lower than 0.2. For example, if exposure and outcome have the same prevalence and the effect of the exposure on the outcome is small, the standardised bias almost equals the unstandardised bias. A standardised bias of 0.10 would then approximately correspond to a bias in estimated risk difference of 10%.

Using equation (2) we plotted the relationship between instrument strength and the threshold sample size for different levels of "standardised bias" in the OLS estimate in Figure 1. For example, for a standardised bias of 0.025-0.05 sample sizes of several

thousand will be needed even if the instrument is very strong (i.e. $\rho_{X,Z.C}$ of 0.4-0.5) to outperform OLS.

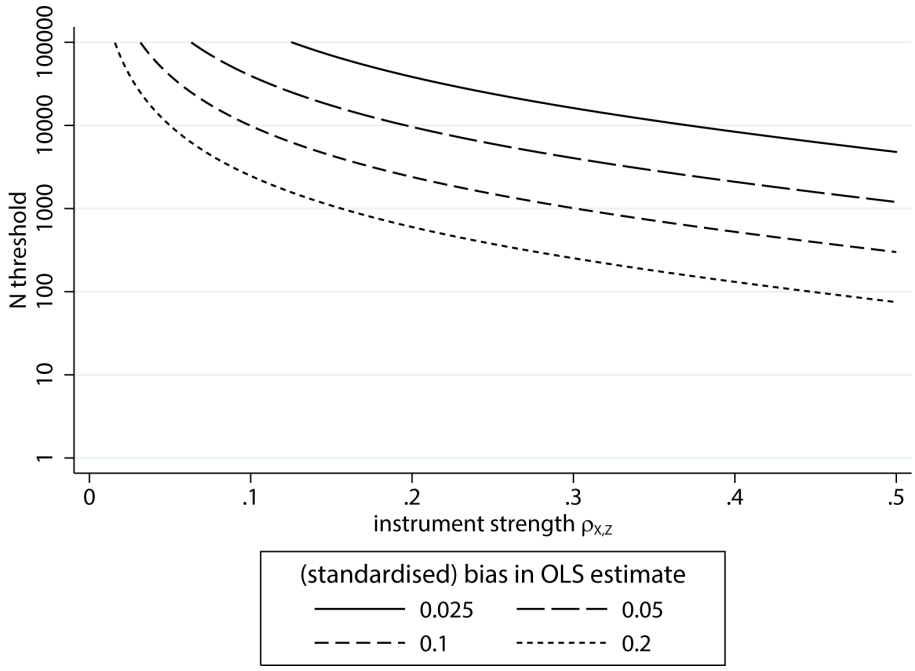


Figure 1. *Plot of the relationship between instrument strength (correlation between instrument Z and exposure X) and the threshold sample size for different levels of 'standardised' bias in the OLS estimate (* bias_{OLS}ó $_{X.C}$/ó $_{Y.X,C}$ *) according to Equation 2.*

'Threshold' sample size: a practical example

To determine how the theoretical considerations above translate to a real application of IV analysis, we applied the formula to the landmark study by Brookhart et al, which was the first to use physician's preference as an IV [7]. Physician's preference may produce variation in treatment prescription unrelated to patient characteristics and prognosis and can therefore be used as an IV. The study used physicians' relative preferences for COX-2 inhibitors in comparison to non-selective NSAIDs as an IV to compare the effect of these drugs on gastrointestinal complications within 120 days

of first prescription [7]. The size of the study population was 49,919 patients [7]. Using multivariable regression a risk difference of -0.06 per 100 patients (95% confidence interval (CI) -0.26 to 0.14) was found. Using IV regression a risk difference of -1.31 per 100 patients (95% CI -2.42 to -0.20) was found. The randomised trial used for comparison with the IV study found a risk difference of -0.65 per 100 patients (95% CI -1.08 to -0.22) [7]. We used the difference between the multivariable regression result and the randomised trial result as the bias of the conventional analysis (0.59/100). We calculated the 'standardised bias', by dividing 0.0059 by $\sqrt{(49{,}919)}\cdot 0.0010$ (i.e. $\sqrt{n}\cdot\hat{\beta}_{OLS}$), because $\mathrm{var}\left(\hat{\beta}_{OLS}\right)=\sigma^2_{Y.X,C}/\left(n\cdot\sigma^2_{X.C}\right)$. This yielded a standardised bias of 0.026. From the confidence intervals we calculated the standard deviation of the OLS risk difference estimate (0.00102) and the IV estimate (0.00566). Because the standard deviation of the IV estimate is $1/\rho_{X,Z.C}$ times larger than the variance of the OLS estimate (see previous section), this means that the correlation between the instrument and the exposure is approximately $\rho_{X,Z.C}$ =0.00102/0.00566 =0.18.

Next, we calculated the threshold sample size above which the IV estimate can be expected to be closer to the true effect than the OLS estimate over a range of levels of bias in the conventional regression estimate, given the correlation between the instrument and the exposure and the value of $\sigma^2_{Y.X,C}/\sigma^2_{X.C}$ as calculated above. The results are displayed in Figure 2. The threshold sample size at the level of bias (0.59 per 100 patients) would be approximately 45,000 patients. This is in the ideal scenario in which all IV assumptions hold and if small sample bias in the two-stage least squares estimator is ignored.

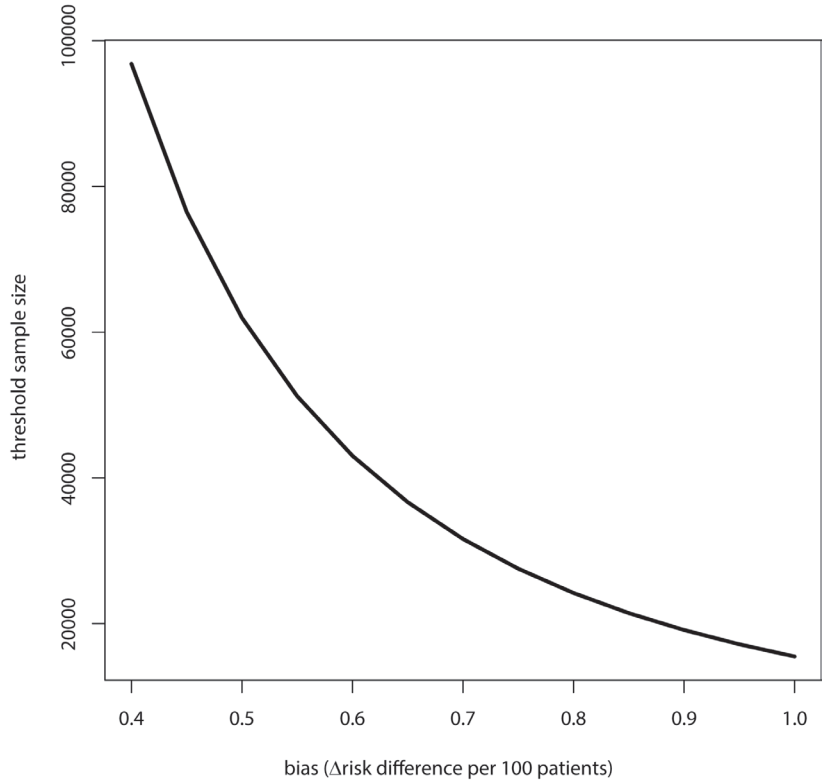


Figure 2. *Example of the relationship between the level of bias of the conventional estimate and the threshold sample size using data from the study by Brookhart. The variance of the IV risk difference estimate was 0.00555², the variance of the OLS estimate was 0.0010² and the study sample size was 32273.*

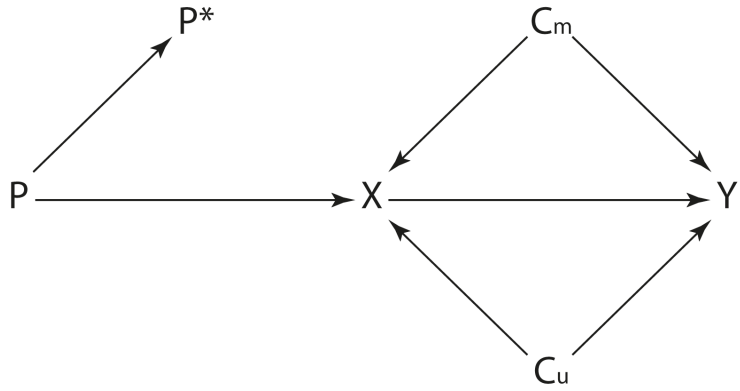


Figure 3. *Directed acyclic graph of the design of the simulation study, with treatment X, outcome Y, measured confounder C_m, unmeasured confounder C_u, physician's preference P and estimated physician's preference P*.*

The role of sample size: simulations

The formula for a threshold sample size calculation discussed above is an approximation. We therefore also performed a simulation study, as this includes the small sample bias of the IV estimator. The design of the simulation study is depicted in a directed acyclic graph in Figure 3. We simulated a hypothetical study investigating the effect of treatment X (X=1: treatment, X=0: no treatment) on continuous outcome Y. We assumed that the study involves patients treated by one of several different physicians and that the probability of receiving treatment X is determined by both patient characteristics (some measured C_m, some unmeasured C_u) and by physician's preference P. The patient characteristics C_m and C_u also affect the outcome Y and are therefore confounders. Physician's preference P is related to treatment, but not to patient characteristics and does not affect outcome Y other than through treatment X, thereby fulfilling IV assumptions. We repeated simulations across a series of sample sizes for four different scenarios regarding instrument strength and strength of unmeasured confounding.

Data generation

We performed Monte Carlo simulations for a series of study population sizes ranging from 500 to 10000 patients (by increments of 500 patients). Per 50 patients a physician with a preference P from the standard uniform distribution U(0,1) was generated (i.e. physician's preference is a continuous variable with a value between 0 and 1). The order of patients in time was defined as the order in which patients were generated. A summary measure of all measured patient characteristics C_m and a summary measure of all unobserved patient characteristic C_u were generated from the uniform distribution U(0,1).

Treatment X was generated from a binomial distribution with individual patients' probabilities of treatment dependent on physician's preference and patient characteristics C_m and C_u according to the following equations:

a. for scenarios 1 and 2 (weaker instrument):

$P(X=1|P,C_m,C_u) = 0.1 + 0.5P + 0.1C_m + 0.2C_u$

This means that two patients with the same values for confounders C_m and C_u would have a 50% difference in probability of receiving treatment if one of these patients had a physician with the highest possible preference (P=1) and the other patient a physician with the lowest possible preference (P=0).

b. for scenarios 3 and 4 (stronger instrument):

$P(X=1|P,C_m,C_u) = 0 + 0.7P + 0.1C_m + 0.2C_u$

I.e. the coefficient for the instrument is larger in the scenarios with the stronger instrument. The intercept is different (0 instead of 0.1), to ensure that the overall probability of receiving treatment is 0.5 in all scenarios.

Continuous outcome Y was generated dependent on treatment X and patient characteristics C_m and C_u as follows:

a. for scenarios 1 and 3 (weaker confounder):

$Y = -1 + 0X + 1C_m + 1C_u + \varepsilon$ with $\varepsilon \sim N(0,1)$

b. for scenarios 2 and 4 (stronger confounder):

$Y = -1.5 + 0X + 1C_m + 2C_u + \varepsilon$ with $\varepsilon \sim N(0,1)$

I.e. the coefficient for the unmeasured confounder is larger in the scenarios with stronger confounding. The intercept was chosen such that the expected value of Y

was 0 in all scenarios. Treatment has no effect on the outcome (indicated by the coefficient of 0).

Estimation of physician preference and data analysis

For a given patient we used the proportion of previous patients of the same physician who received treatment *X* as the estimated preference of his physician (indicated by *P** in the directed acyclic graph of Figure 3) for use in the IV analysis (as true preference *P* would not be known). The IV used in the original description of physician preference-based IV analysis [7], the treatment of only the last previous patient, would be a much weaker instrument in these simulations (in which preference *P* does not change over time), because it is not as strongly related to *P* as an instrument based on multiple previous prescriptions. The treatment effect was estimated using OLS and 2-SLS IV regression, both adjusted for measured patient characteristic C_m.

Monte Carlo simulations and subsequent analyses were performed using Stata version 12 (College Station, TX: StataCorp LP. 2011).

Simulation results

The first stage regression results, including the partial r^2 and the first stage F-statistic are discussed in eAppendix 2 and eTable 1. The mean squared errors (graphically displayed as root MSEs) of the different analyses across the range of sample sizes in the four different scenarios are depicted in Figure 4. We excluded the samples in which the first stage F-statistic was below 10 (to exclude those samples in which an IV analysis would be unlikely to be performed). In scenario 1 (weaker instrument, weaker confounding), the MSE of the 2-SLS estimates was much larger than that of

the OLS estimate at smaller sample sizes. Whereas the MSE of the OLS estimate only decreased minimally with increasing sample size, the MSE of the 2-SLS estimates decreased substantially, although the 2-SLS estimates were still further on average from the true effect than the OLS estimates at sample size 10,000 (in further simulations for this scenario the 2-SLS estimates outperformed the OLS estimates in terms of MSE from 25,000 patients upwards).

Scenario 2 (weaker instrument, stronger confounding) resulted in a larger MSE for the OLS estimates. These more biased OLS estimates were first outperformed by the 2-SLS estimates at a smaller sample size of 7000 patients. The stronger instrument in

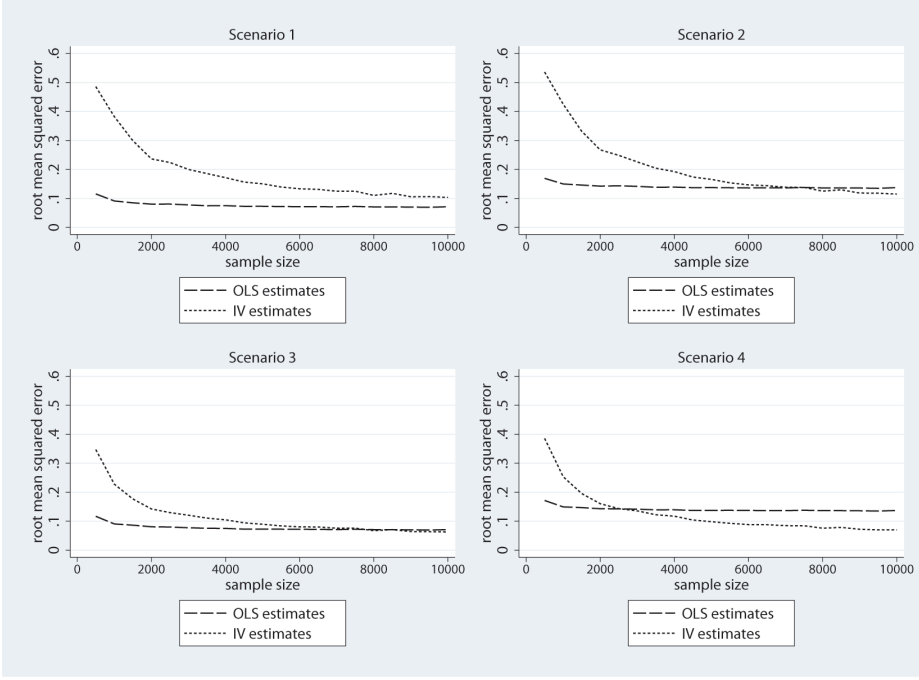


Figure 4. *Plots of the root mean squared errors of the linear regression estimates and IV regression estimates in the following simulation scenario's:*
Scenario 1: weaker instrument, weaker unmeasured confounding.
Scenario 2: weaker instrument, stronger unmeasured confounding.
Scenario 3: stronger instrument, weaker unmeasured confounding.
Scenario 4: stronger instrument, stronger unmeasured confounding.

scenario 3 (stronger instrument, weaker confounding) resulted in a reduced variance of 2-SLS estimates, closer on average to true effect than the OLS estimates at sample sizes of 8000 upwards. Due to the combination of a stronger instrument and stronger confounding the sample size at which the 2-SLS estimates were on average closer to the true effect than the OLS estimates by sample size 3000 in scenario 4.

In our simulations the threshold value for scenario 4 was approximately 2600 patients (Figure 2). Using equation 2, we calculated what the approximate threshold sample size would be, given the bias of the OLS estimates in scenario 4 (0.133) and the partial r^2 (0.111) (See eTable 1). This would give a threshold sample size of about 2400 (See eAppendix 3), i.e. very similar to the threshold value observed in our simulation results, indicating that equation 2 gives a good approximation.

Discussion

We have shown how the performance of IV analysis in comparison to conventional analyses depends substantially on sample size. We have provided an equation that can be used to approximate a 'threshold' sample size above which the mean squared error of IV analyses will be lower than that of conventional analyses in case IV assumptions hold. We show that substantial sample sizes will generally be needed for IV analyses to provide a better estimate on average than conventional analyses in epidemiologic studies.

In smaller studies the IV estimate may have a wide, uninformative, confidence interval, a point estimate far from the true value, and the possibility of small sample

bias. Although the average deviation of the OLS from the true value is smaller, the coverage of the 95% confidence will be generally smaller than 95%, because of the bias in the estimate. In practice, the direction of the bias based on OLS estimates is often known, especially in case of confounding by indication. Therefore we recommend in small studies to calculate the OLS estimate, accompanied by sensitivity analysis for bias due to unmeasured confounding as described by for example Greenland [21], rather than performing an IV analysis.

For the present study we chose the MSE as a measure of performance of the different analyses, rather than the coverage of the confidence interval, because the MSE incorporates both the bias and the variance of estimates whereas the coverage of the confidence interval provides information on the type I error rate alone. Although we would not recommend in general to prefer the method with the smallest MSE, it should be considered if the trade-off is between bias with known direction and estimable magnitude and an extremely large variance. Further, the size of the point estimate inevitably draws the attention of the researcher and reader, which means that its deviance from the true effect is therefore also important (alongside correct coverage of the confidence interval). Another reason why the large variance of IV analysis should be noted is that publication bias of statistically significant findings would, in the case of the very large variances seen in IV analysis, lead to the publication of very large effect estimates.

The trade-off between the bias of conventional estimates and the variance of IV estimates has been described and discussed previously, but not with a focus on the role of sample size. One simulation study varied the proportion of patients whose treatment was affected by the physician's preference, thereby varying the strength of

the instrument and the variance of the IV estimates, but did not vary the sample size [12]. A health econometrics study compared IV estimates and OLS estimates varying the degree of violation of IV assumptions, instrument strength and sample size, all in case of extremely biased OLS estimates [13]. Here we have shown how sample size in itself can limit the usefulness of IV analysis in comparison to conventional analyses in case of moderately to considerably biased OLS estimates.

A limitation of the equation we use for calculation of a threshold sample size is that it does not take into account the small sample bias (weak instrument bias) of the IV estimate. The level of small sample bias depends on the strength of the instrument and the sample size and is indicated by the first-stage F-statistic; an F-statistic above 10 is generally considered sufficient for small sample bias to be negligible [11]. If the instrument is weak and small sample bias likely to be present, the benefit of IV analysis over OLS analysis will be further reduced. As our equation is intended to compare 2-SLS and OLS in terms of mean squared error, it cannot be used as a formal power calculation. Power calculations for IV analysis have been described by others [15;22].

The trade-off between bias of conventional analysis methods and variance of IV methods is a general issue and will also apply for alternative IV approaches such as 2-stage logistic models [23], and similar types of calculations should be possible.

We applied our threshold sample size equation to an example of an IV study by Brookhart et al [7], which yielded a very large threshold sample size of approximately 45,000 patients. Moreover, estimates from a physician preference based IV analysis that we performed in a study population of 476 patients had a very

large variance, despite a strong instrument (in press). This indicates the large variance of IV estimates at sample sizes of this order forms a problem in real applications of IV methods and not just in the particular design of our simulations.

IV analysis may be useful in studies involving either a very strong IV *or* a very large sample. Typical clinical observational studies are unlikely to meet either of these conditions. However, IV analysis may also be used to adjust for incomplete compliance in randomised controlled trials. The instrument is then the randomisation, which plausibly meets IV assumptions and will generally be strongly related to the treatment received (unless compliance is very low). The IV analysis will then estimate the treatment effect among the compliers (assuming there are no defiers, i.e. no patients who would always receive the opposite of the treatment they are randomised for). Other opportunities for useful application of IV analysis will increasingly become available as we enter the era of "big data". Within increasingly available very large cohorts IV methods can be feasible and useful; they will have an advantage over other methods if information on confounders is insufficient.

In conclusion, even when all IV assumptions hold, IV analysis can only be expected to provide an estimate closer to the true treatment effect than conventional estimates in large study populations. The size of the study population above which this is the case depends on the strength of the available instrument and the expected amount of unmeasured confounding. IV methods will therefore be of most value in large studies of intended effects, when considerable unmeasured confounding is likely, but only if a strong and plausible instrument is available.

Acknowledgements

Sources of funding

This work was supported by the Netherlands Organisation for Health Research and Development (ZonMw, grant number 152002040).

Conflicts of interest

The authors declare no conflict of interest.

References

(1) Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiol Drug Saf 2010 Jun;19(6):537-54.

(2) Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. J Clin Epidemiol 2009 Dec;62(12):1226-32.

(3) Swanson SA, Hernan MA. Commentary: how to report instrumental variable analyses (suggestions welcome). Epidemiology 2013 May;24(3):370-4.

(4) Davies NM, Davey Smith G, Windmeijer F, Martin RM. Issues in the reporting and conduct of instrumental variable studies: a systematic review. Epidemiology 2013 May;24(3):363-9.

(5) Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiology 2006 Jul;17(4):360-72.

(6) Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA 2007 Jan 17;297(3):278-85.

(7) Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. Epidemiology 2006 May;17(3):268-75.

(8) Schneeweiss S, Seeger JD, Landon J, Walker AM. Aprotinin during coronary-artery bypass grafting and risk of death. N Engl J Med 2008 Feb 21;358(8):771-83.

(9) Schneeweiss S, Setoguchi S, Brookhart A, Dormuth C, Wang PS. Risk of death associated with the use of conventional versus atypical antipsychotic drugs among elderly patients. CMAJ 2007 Feb 27;176(5):627-32.

(10) Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med 2008 Apr 15;27(8):1133-63.

(11) Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. Epidemiology 2006 May;17(3):260-7.

(12) Ionescu-Ittu R, Delaney JA, Abrahamowicz M. Bias-variance trade-off in pharmacoepidemiological studies using physician-preference-based instrumental variables: a simulation study. Pharmacoepidemiol Drug Saf 2009 Jul;18(7):562-71.

(13) Crown WH, Henk HJ, Vanness DJ. Some cautions on the use of instrumental variables estimators in outcomes research: how bias in instrumental variables estimators is affected by instrument strength, instrument contamination, and sample size. Value Health 2011 Dec;14(8):1078-84.

(14) Abrahamowicz M, Beauchamp ME, Ionescu-Ittu R, Delaney JA, Pilote L. Reducing the variance of the prescribing preference-based instrumental variable estimates of the treatment effect. Am J Epidemiol 2011 Aug 15;174(4):494-502.

(15) Freeman G, Cowling BJ, Schooling CM. Power and sample size calculations for Mendelian randomization studies using one genetic instrument. Int J Epidemiol 2013 Aug;42(4):1157-63.

(16) Burgess S, Thompson SG. Improving bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. Stat Med 2012 Jul 10;31(15):1582-600.
(17) Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables II: instrumental variable application-in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. J Clin Epidemiol 2009 Dec;62(12):1233-41.
(18) Huybrechts KF, Brookhart MA, Rothman KJ, Silliman RA, Gerhard T, Crystal S, et al. Comparison of different approaches to confounding adjustment in a study on the association of antipsychotic medication with mortality in older nursing home patients. Am J Epidemiol 2011 Nov 1;174(9):1089-99.
(19) Hadley J, Yabroff KR, Barrett MJ, Penson DF, Saigal CS, Potosky AL. Comparative effectiveness of prostate cancer treatments: evaluating statistical adjustments for confounding in observational data. J Natl Cancer Inst 2010 Dec 1;102(23):1780-93.
(20) Rassen JA, Mittleman MA, Glynn RJ, Alan BM, Schneeweiss S. Safety and effectiveness of bivalirudin in routine care of patients undergoing percutaneous coronary intervention. Eur Heart J 2010 Mar;31(5):561-72.
(21) Greenland S. Basic methods for sensitivity analysis of biases. Int J Epidemiol 1996 Dec;25(6):1107-16.
(22) Brion MJ, Shakhbazov K, Visscher PM. Calculating statistical power in Mendelian randomization studies. Int J Epidemiol 2013 Oct;42(5):1497-501.
(23) Rassen JA, Schneeweiss S, Glynn RJ, Mittleman MA, Brookhart MA. Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. Am J Epidemiol 2009 Feb 1;169(3):273-84.

eAppendix 1: Derivation of equation 2

We assume the following underlying true model:

$E[Y|X, C, U] = \alpha + \beta X + \gamma C + \delta U$,

with C a vector of known confounders and U a vector of unknown confounders.

Our aim was to determine from which sample size IV estimates would be on average closer to the true effect than OLS estimates. We assume that we have data of a study of size n, in which we have estimated the true treatment effect β by an IV regression analysis, yielding estimate $\hat{\beta}_n^{IV}$ and by OLS, yielding estimate $\hat{\beta}_n^{OLS}$, adjusting in both analyses for known confounders.

In the situation without confounders the two step IV estimator equals the ratio of the sample covariances:[1;2]

$$\hat{\beta}_n^{IV} = \hat{\sigma}_{Z,Y}/\hat{\sigma}_{Z,X}.$$

The asymptotic variance of this estimator is

$$\text{var}(\hat{\beta}_n^{IV}) = \frac{\sigma_{Y.X}^2}{n\sigma_X^2\rho_{X,Z}^2},$$

with $\sigma_{Y.X}^2$ the residual variance of Y after adjusting for X, σ_X^2 the variance of the exposure X and $\rho_{X,Z}$ the correlation between the exposure X and instrument Z, which reflects the strength of the instrument.[1;3]

The variance of the OLS estimate is $\text{var}\left[\hat{\beta}_n^{OLS}\right] = \frac{\sigma_{Y.X}^2}{\sigma_X^2 n}$

As a measure for the average deviation from the true effect we use the mean squared error (MSE), where the MSE of an estimate $\hat{\beta}_n$ is defined as

$$\mathrm{MSE}\left(\hat{\beta}_n\right) = \mathrm{E}\left[\left(\hat{\beta}_n - \beta\right)^2\right]$$

$$= \mathrm{E}\left[\left(\hat{\beta}_n - \mathrm{E}\left[\hat{\beta}_n\right]\right)^2\right] + \left(\mathrm{E}\left[\hat{\beta}_n\right] - \beta\right)^2$$

$$= \mathrm{var}\left[\hat{\beta}_n\right] + \mathrm{bias}\left[\hat{\beta}_n\right]^2$$

In the best-case scenario in which IV assumptions hold and small sample bias can be ignored, the bias of the IV estimator would be zero. Then

$$\mathrm{MSE}\left(\hat{\beta}_n^{\mathrm{IV}}\right) = \mathrm{var}\left[\hat{\beta}_n^{\mathrm{IV}}\right]$$

The bias of the OLS estimate does not depend on the sample size n. Hence

$$\mathrm{MSE}\left(\hat{\beta}_n^{\mathrm{OLS}}\right) = \mathrm{var}\left[\hat{\beta}_n^{\mathrm{OLS}}\right] + \mathrm{bias}_{\mathrm{OLS}}^2$$

For the threshold sample size n_t, the $\mathrm{MSE}_{\mathrm{IV}} = \mathrm{MSE}_{\mathrm{OLS}}$ and

$$\mathrm{var}\left[\hat{\beta}_{nt}^{\mathrm{IV}}\right] = \mathrm{var}\left[\hat{\beta}_{nt}^{\mathrm{OLS}}\right] + \mathrm{bias}_{\mathrm{OLS}}^2$$

Filling in the formulas for the variance of the OLS and the IV estimate we obtain that for the threshold sample size n_t:

$$\frac{\sigma_{Y.X}^2}{n_t \sigma_X^2 \rho_{x,z}^2} = \frac{\sigma_{Y.X}^2}{n_t \sigma_X^2} + \mathrm{bias}_{\mathrm{OLS}}^2,$$

from which it follows that

$$n_t = \frac{\sigma_{Y.X}^2}{\sigma_X^2 \mathrm{bias}_{\mathrm{OLS}}^2}\left(\frac{1}{\rho_{X,Z}^2} - 1\right).$$

Extension to the situation with adjustments for measured confounders C is straightforward, by controlling variances and correlations for C, i.e σ_X^2 is replaced by $\sigma_{X.C}^2$ the residual variance of X after adjusting for C in a linear regression analysis, $\sigma_{Y.X}^2$ is replaced by $\sigma_{Y.X,C}^2$, and $\rho_{X,Z}$ is replaced by $\rho_{X,Z.C}$ the partial correlation between X and Z, adjusted for C.

References

[1] Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. *Epidemiology* 2006;17:260-267.
[2] Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 1996;91:444-455.
[3] Freeman G, Cowling BJ, Schooling CM. Power and sample size calculations for Mendelian randomization studies using one genetic instrument. *Int J Epidemiol* 2013;42:1157-1163.

eAppendix 2: Strength of the instrument in the simulations

eTable 1 displays the mean F-statistic and partial r^2 for the first-stage regression across the range of sample sizes for the weaker instrument (scenario 1&2) and the stronger instrument (scenario 3&4). The strength of the instrument-outcome association varied from one sample to the next, which resulted in some samples with a very weak instrument by chance at the smallest sample sizes. For the weaker instrument, 29.7% of samples had an F-statistic smaller than 10 at sample size 500. This quickly decreased with increasing sample size, all samples having an F-statistic above 10 at sample size 2000. For the stronger instrument 3.6% of samples had an F-statistic smaller than 10 at sample size 500 and none at a larger sample size. The average partial r^2 was 0.041 for the weaker instrument and 0.111 for the stronger instrument (at sample size 10000).

eTable 1: *First-stage regression statistics of the simulation study*

n	Scenario 1&2			Scenario 3&4		
	Average F-statistic	F-statistics <10 (%)	Average partial r^2	Average F-statistic	F-statistics <10 (%)	Average partial r^2
500	19.0	29.7	0.037	55.1	3.6	0.099
1000	40.5	4.8	0.039	117.9	0	0.106
1500	61.6	0.2	0.040	178.8	0	0.108
2000	83.1	0	0.041	241.4	0	0.109
2500	102.7	0	0.040	298.2	0	0.108
3500	148.4	0	0.041	427.9	0	0.111
4000	169.6	0	0.041	488.5	0	0.110
4500	187.3	0	0.041	544.8	0	0.110
5000	209.5	0	0.041	609.5	0	0.110
5500	234.2	0	0.042	676.9	0	0.111
6000	252.4	0	0.041	729.9	0	0.110
6500	276.4	0	0.042	793.7	0	0.111
7000	300.4	0	0.042	864.1	0	0.112
7500	314.2	0	0.041	911.2	0	0.110
8000	337.3	0	0.041	977.5	0	0.111
8500	360.4	0	0.041	1043	0	0.111
9000	380.8	0	0.041	1102	0	0.111
9500	403.2	0	0.041	1166	0	0.111
10000	422.5	0	0.041	1221	0	0.111

Average first-stage F-statistic, % of first-stage F-statistics <10 and average first stage partial r2 across 1000 samples at each sample size for the weaker instrument (scenario 1&2) and the stronger instrument (scenario 3 and 4).

eAppendix 3: Threshold sample size calculation for simulation scenario 4.

In this scenario the threshold sample size would be

$$n_{\text{threshold}} = \frac{\sigma^2_{Y.X,C_m}}{\sigma^2_{X.C_m}\,\text{bias}^2_{OLS}}\left(\frac{1}{\rho^2_{X,Z.C_m}} - 1\right),$$

with Z the instrument used .

In scenario 4 we simulated from the following model:

$P(X=1|P,C_m,C_u) = 0 + 0.7P + 0.1C_m + 0.2C_u$ and

$Y = -1.5 + 0X + 1C_m + 2C_u + \varepsilon$ with $\varepsilon \sim N(0,1)$,

with C_m, C_u,P ~uniform $(0,1)$.

In this scenario $\sigma^2_{Y.X,C_m} = 4\text{var}(C_u) + \text{var}(\varepsilon)$.

The variance of the unknown confounder is equal to var $(C_u) = 1/12$, because C_u is uniformly distributed on $[0,1]$, which gives $\sigma^2_{Y.X,C_m} = 1.33$.

The residual variance of X after controlling for C_m is $\sigma^2_{X.Cm} = (1-\rho^2_{X.Cm})\sigma^2_X$.

Since $\rho^2_{X,C} = 0.1^2\sigma^2_{Cu}/\sigma^2_X = 0.0033$, $\sigma^2_{X.Cm} = (1-0.0033)\,0.25 = 0.25$.

In the simulations the mean bias of the OLS estimates in scenario 4 was 0.133 and the mean partial r^2 between the instrument Z and treatment X was 0.111.

Using this yields

$$n_{\text{threshold}} = \frac{1.33}{0.25 \cdot 0.133^2}\left(\frac{1}{0.111} - 1\right) = 2415$$

Chapter 5

Instrumental variable analysis as a sensitivity analysis in studies of adverse effects: venous thromboembolism and 2nd vs. 3rd generation oral contraceptives.

Anna G.C. Boef, Patrick C. Souverein, Jan P. Vandenbroucke, Astrid van Hylckama Vlieg, Anthonius de Boer, Saskia le Cessie and Olaf M. Dekkers

Submitted

Abstract

Purpose: A potentially useful role for instrumental variable (IV) analysis may be as a sensitivity analysis to assess the presence of confounding when studying adverse drug effects. There has been discussion on whether the observed increased risk of venous thromboembolism (VTE) for 3rd generation oral contraceptives versus 2nd generation oral contraceptives could be (partially) attributed to confounding. We investigated how prescribing preference IV estimates compare to conventional estimates.

Methods: Women in the Clinical Practice Research Database who started a 2nd or 3rd generation oral contraceptive from 1989-2013 were included. Ordinary least squares and two-stage least squares regression were used to estimate risk differences in VTE. Cox regression and IV for Cox proportional hazards regression were used to calculate hazard ratios (HR). The instrument used was the proportion of prescriptions for 3rd generation oral contraceptives by the general practitioner in the year preceding the current prescription.

Results: All analyses pointed in the direction of an increased VTE risk for 3rd generation oral contraceptives. The adjusted HR from the conventional Cox regression was 1.62 (95% CI 1.16-2.27) and the fully adjusted HR from the IV Cox regression was 3.45 (95% CI 0.97-11.7), showing a larger risk and wider confidence intervals in the IV analysis.

Conclusions: The similarity in direction of results from the IV analyses and conventional analyses suggests major confounding is unlikely. The conventional observational analysis may have been conservative. IV analysis can be a useful sensitivity analysis to assess the presence of confounding in studies of adverse drug effects in very large databases.

Introduction

Observational data analyses of intended effects of drug therapy are always suspected to be strongly confounded by factors that determine prognosis. This has been termed 'confounding by indication'.[1] However, it has been argued, and there is empirical evidence, that this is not the case for adverse effects.[2-4] Although confounding by contra-indication[5] can exist, for many adverse effects of treatments little confounding is expected because these adverse effects are difficult to predict.[2;6] Still, controversies can emerge due to different views about the potential for confounding when studying adverse effects. Performing an instrumental variable (IV) analysis may then be a consideration, because this method rests upon different assumptions. It requires identification of a variable that determines treatment but is not otherwise associated with the outcome - thereby mimicking randomisation. We explore the value of IV analysis as a 'sensitivity analysis' to assess the presence of confounding when studying adverse effects.

As an example we use the controversy about the risk of venous thromboembolism (VTE) of 3rd vs 2nd generation combined hormonal oral contraceptives (OCs). In general, it can be expected that prescribers did not take a patient's thrombosis risk into account when choosing between different OCs before 1995, when evidence of an increased risk of VTE in 3rd generation in comparison to 2nd generation OCs was published.[2;7-9] Users of different classes of OCs before 1995, included in the studies published in 1995 and major studies based on data from before 1995,[7-11] can therefore be expected to have had a comparable background risk of VTE. However, there was an extensive debate on (the direction of) the relation between thrombosis risk and OC choice in these studies and the resulting confounding.[12-14] After 1995 general practitioners (GPs) will have become aware of the increased VTE risk for 3rd generation OCs and may have started taking patients' thrombosis risk into account when choosing between OCs.[7-9] Yet this risk is difficult to predict in young women, and we therefore expect confounding by contra-indication to have remained limited.

If the observed difference in risk of VTE between 2nd and 3rd generation OCs were only based on confounding, in principle, an IV analysis (e.g. using GP's preference as an instrument) should show no difference in VTE incidence. On the other hand, if there is indeed little confounding by contra-indication for the association between 3rd vs 2nd generation OCs and VTE, effect estimates from conventional analyses and IV analyses should yield similar results. Therefore, we investigated how GP's preference IV estimates of the effect of 3rd vs 2nd generation OCs on occurrence of VTE compare to conventional estimates from observational data.

Methods

Study population

The study population comprised all women aged 15-44 with a first prescription for a combined hormonal OC between 1987 and 2013 included in the Clinical Practice Research Datalink (CPRD). Those with a first prescription within six months of their registration date or date on which practice data were up-to-standard were excluded (n=366 354) as this may be a repeat prescription. Further reasons for exclusion were: a prescription for emergency contraceptives only (n=11 575), a first prescription with a repeat prescription code (n=29), occurrence of VTE before the first prescription (n=509) or an unknown prescriber for the first prescription (n=11561). Of the 502163 remaining women, 444542 were first prescribed a 2nd or 3rd generation combined hormonal OC (as defined below) and were included in the study population.

Exposure

Second generation OCs were defined as OCs containing levonorgestrel, lynestrenol or norethisterone as a progestagen and < 50μg of oestrogen. Third generation OCs were defined as OCs containing desogestrel, gestodene or norgestimate progestagen and < 50μg of oestrogen. Supplementary Codelist 1 lists the codes used. Users of contraceptives containing other progestagens, such as drospirenone, were not included.

Outcomes

Outcomes were defined based on records of Read codes for deep vein thrombosis or pulmonary embolism (Supplementary Codelist 2). Codes for deep vein phlebitis or thrombophlebitis were also included, as these may contain misclassified VTE events. An additional requirement was prescription of anticoagulant treatment in the period from 1 month before until 6 months after the diagnosis code date. For the analyses estimating risk differences we included events which occurred within one year or three years after the first OC prescription. Patients who started OCs after 31st December 2012 (1 year) or after 31st December 2010 (3 years) were excluded from these analyses. For the Cox regression analyses only events occurring during the continuous period of use of the same OC since the first prescription were included.

Other patient characteristics

Information on smoking and BMI was obtained if available. The most recent information before the first prescription for OCs was used, with a minimum age of 12 for smoking behaviour and 14 for BMI. BMI values >14 or <60 were excluded.

Instrument definition

We used previous prescriptions of the patient's GP as a proxy for GP's preference for 2nd or 3rd generation OCs at the time of that patient's first OC prescription. We considered the following instruments:

1. The previous first-time OC prescription of the same GP.
2. The proportion of 3rd generation OCs among the previous five first-time prescriptions of the same GP.
3. The proportion of 3rd generation OCs among all first-time prescriptions of the same GP in the year preceding the current treatment decision.

The strength of all three instruments (first-stage risk difference for 3rd generation OC prescription, partial r^2 and partial F-statistic) was determined (Supplementary Table 1). Instrument 3 (unadjusted: partial r^2 0.191, F-statistic 99357; adjusted for calendar time: partial r^2 0.085, F-statistic 38881) was selected for use in the IV analyses. We preferred this instrument over instrument 2 because of the fixed time interval in which preference was determined. All further analyses were performed in the 420152 subjects with a value for this instrument (excluding subjects whose GP had not prescribed any combined hormonal OC in the year preceding their prescription date).

IV assumptions

Figure 1 depicts the assumed causal relations in this study. For previous prescription(s) of the GP to be a valid instrument the following assumptions must hold:

1. Previous first-time prescriptions of the same GP for 2nd or 3rd generation OCs must be associated with the type of OC prescribed to the current patient.
2. The prescriptions of previous patients may not affect the VTE risk of the current patient other than through the type of OC the current patient receives.
3. The prescriptions of previous patients and the baseline VTE risk of the current patient do not have a common cause.

In order to obtain a point estimate we further assume stochastic monotonicity, under which a strength-of-IV weighted average treatment effect is estimated (for details we refer to elsewhere).[15]

Statistical analyses

All statistical analyses were performed using Stata 10.1.

1. Ordinary least squares analysis (OLS)
   First OLS regression was used to estimate the difference in risk of VTE between users of 3rd generation OCs and users of 2nd generation OCs 1 year and 3 years after first prescription. We performed these analyses: 1) unadjusted; and 2) adjusted for calendar year (using year continuously, ≤1995 versus ≥1996, and their interaction term), age, BMI and smoking (non-smoker/ smoker/ ex-smoker).

2. Two-stage least squares analysis (2-SLS)
   Next 2-SLS regression was performed, using the instrument selected previously. The estimates obtained are risk differences of VTE for 3rd vs 2nd generation OCs. Heteroscedasticity-robust standard errors were used. We performed these analyses: 1) unadjusted; 2) adjusted for calendar year; and 3) adjusted for calendar year, age, BMI and smoking.
3. Cox proportional hazards regression
   Next, Cox proportional hazards regression was performed to estimate the hazard ratio for venous thrombolism for users of 3rd generation OCs versus users of 2nd generation OCs. We used the full period of uninterrupted use of the first prescribed OC as the observation period (ending if OC use was stopped, if a switch to another class of OC was made (e.g. from a 2nd to a 3rd generation OC), if the patient was no longer registered in the practice, if the patient developed a VTE, or if the patient died). We performed this analysis: 1) unadjusted; and 2) adjusted for calendar year, age, BMI and smoking.
4. IV for Cox proportional hazards regression
   We used an adapted version of IV regression to take into account the length of follow-up. The model used was an IV for Cox proportional hazards model, the use of which has been shown to be appropriate in case of a rare outcome.[16] The first stage of this model is linear regression of the treatment on the instrument (and,

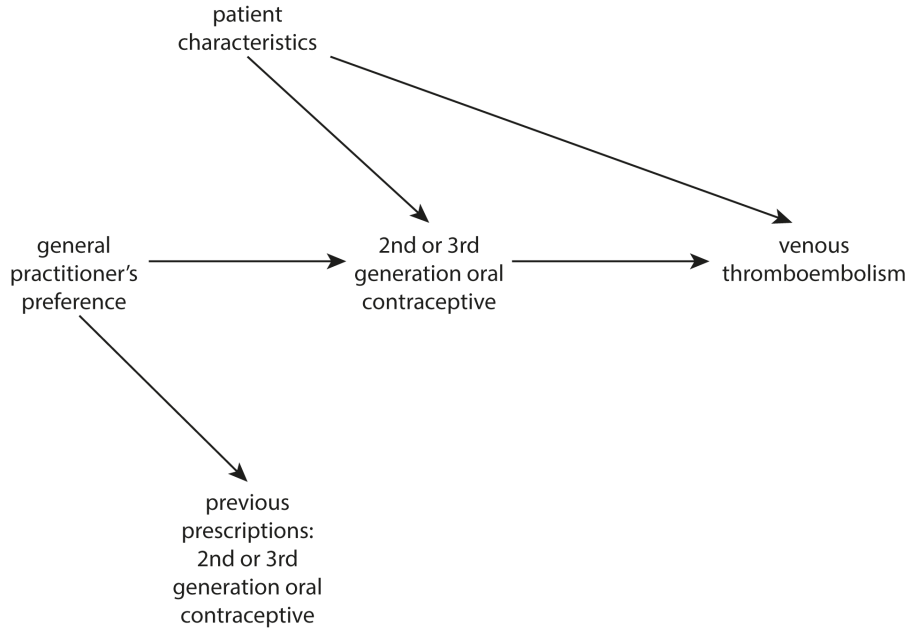


Figure 1. *Directed acyclic graph of the assumed causal relations in this study.*

for the adjusted analysis, the covariates). The second stage is Cox regression, with the fitted probability of a 3rd generation OC from the first stage as the independent variable (and, for the adjusted analysis, including the same covariates as in the first stage). To obtain a confidence interval we used nonparametric bootstrap (1000 runs). We performed the analysis: 1) unadjusted; 2) adjusted for calendar year; and 3) adjusted for calendar year, age, BMI and smoking. For the fully adjusted analysis the average of the 2.5th and 97.5th bootstrap percentile across the 10 imputations were used as an approximation, which gives a slightly too narrow confidence interval.

Initially, we planned to perform all analyses in two time periods, namely the time periods before and after publication of evidence of an increased risk of VTE for third generation OCs in 1995; i.e. 1987-1994 and 1996-2011. Unfortunately, due to the low number of patients newly starting a 2nd or 3rd generation OC before 1995 (n=46747, n=45354 with a value of the instrument), this was not feasible.

Missing values

Missing values for BMI and smoking were imputed using multiple imputation using chained equations, using linear regression for BMI and multinomial logistic regression for smoking. All versions of the outcome from the different analyses, log-transformed follow-up time, the exposure (second or third generation OCs), the instrument and all covariates were included in the imputation model.

Sensitivity analyses

The exclusion of women with a first prescription within six months of the entry date may not be sufficient to exclude all patients for whom the first prescription recorded is a repeat prescription. We therefore performed sensitivity analyses in which we excluded all patients with a first prescription within a year of the entry date.

The requirement of a record of a prescription for anticoagulation by the GP within 6 months after the potential thromboembolic event may be too strict, as some patient may have received all prescriptions for anticoagulants via the hospital. We therefore performed sensitivity without this requirement.

Results

Patient characteristics

Characteristics of the study subjects are shown in Table 1, both by actual treatment and by value of the instrument. Patients who received a 3rd generation OC were older (median age 24.3 years versus 20.3 years) and smoked slightly more (26.5% versus 25.0%) than patients who received a 2nd generation OC. As the percentage of 3rd generation prescriptions in the preceding year by the same GP increased, the age of the patients increased (median of 23.2 years in the highest group versus 20.0 years in the lowest group) and the percentage of smokers also increased (highest group: 27.3%; lowest group: 23.8%).

Table 1. *Patient characteristics by actual type of oral contraceptive and by quintiles of prescriptions for 3rd generation oral contraceptives by their GP in the past year.*

	Actual prescription		Prescriptions of the same GP in past year (% 3rd generation)			
	2nd generation	3rd generation	0	Q1 (1.6-20.0)	Q2 (20.2-44.4)	Q3 (44.6-100)
N	309508	110644	133349	98423	94119	94261
Actual prescription 3rd generation	N/A	N/A	16121 (12.1)	15068 (15.3)	23967 (25.5)	54796 (58.7)
Age (y), median (IQR)	20.3 (17.0-28.1)	24.3 (18.5-30.6)	20.0 (17.0-27.7)	21.0 (17.2-28.7)	22.1 (17.6-29.4)	23.2 (17.9-29.9)
BMI (kg/m²), median (IQR)†	22.9 (20.6-26.2)	22.8 (20.7-25.8)	23.0 (20.5-26.3)	23.0 (20.7-26.4)	22.9 (20.6-26.0)	22.7 (20.6-25.7)
Smoking‡						
Yes	62515 (25.0)	22612 (26.5)	26356 (23.8)	20408 (25.4)	19414 (25.9)	18949 (27.3)
Ex	20304 (8.1)	7800 (9.2)	9349 (8.4)	6997 (8.7)	6470 (8.6)	5288 (7.6)
Venous thrombo-embolism*						
1 year	91 (0.03)	45 (0.04)	48 (0.04)	27 (0.03)	26 (0.03)	35 (0.04)
3 years	180 (0.07)	101 (0.10)	81 (0.08)	55 (0.06)	64 (0.08)	81 (0.09)

Data are presented as n(%) unless stated otherwise.

†Data available for 239593 patients.

‡Data available for 335553 patients.

*Available for 403864 patients (1year) and 364211 patients (3 years).

Changes in prescription behaviour over time

A reason why the instrument was related to age and smoking is that the instrument was strongly related to calendar time. In Figure 2, we show the proportion of prescriptions for 3rd generation and 2nd generation OCs (and drospirenone-containing contraceptives) per calendar year. From 1989 (40%) to 1994 (78%) the proportion of prescriptions for 3rd generation OCs increased. During 1995 (68%) this trend stopped, leading into a drop in 3rd generation OC prescriptions in 1996 (21%). After 1996 the proportion of 2nd generation OCs remained relatively constant between 75%-80%, with the proportion of 3rd generation OCs gradually decreasing as the proportion of drospirenone-containing OCs increased. Supplementary Table 7 shows that the age at first prescription and the proportion of smokers decreased over time.

OLS and 2-SLS regression

Differences in 1-year and 3-year risk of VTE between 3rd generation and 2nd generation OCs obtained using OLS regression and 2-SLS IV regression are displayed in Table 2. All OLS results show an increased risk for VTE for 3rd generation OCs in comparison to 2nd generation OCs: the adjusted 1-year risk difference was 1.2 events per 10000 patients (95% CI -0.2;2.6) and the adjusted 3-year risk difference was 2.0 events per

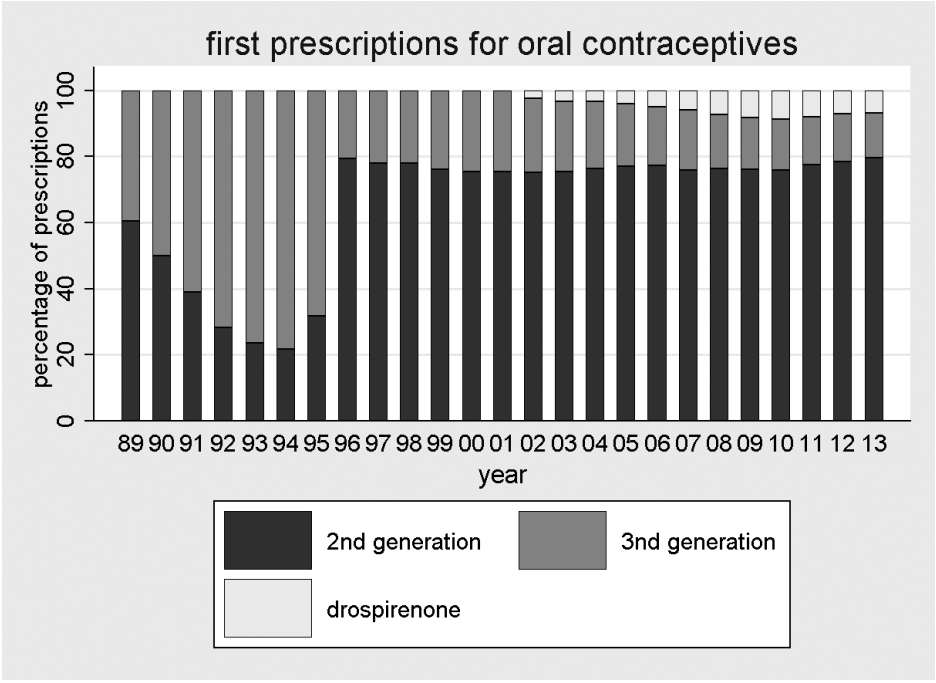


Figure 2. *Proportion of prescriptions for 3rd generation and 2nd generation oral contraceptives (and drospirenone-containing contraceptives) per calendar year.*

10000 patients (-0.2;4.2). All point estimates from the 2-SLS regression were also in the direction of an increased risk for VTE for 3rd generation OCs, but with much wider confidence intervals. The fully adjusted 2-SLS analyses gave a 1-year risk difference of 3.5 events per 10000 patients (-1.2;8.3) and a 3-year risk difference of 3.0 events per 10000 patients (-4.5;10.4).

Table 2. *Conventional and instrumental variable estimates of the risk differences of venous thromboembolism per 10,000 patients for 3rd generation versus 2nd generation oral contraceptives within 1 year and within 3 years of first prescription.*

		Ordinary least squares (Risk difference per 10,000)		Two-stage least squares (Risk difference per 10,000)		
Time	No. Events	Unadjusted	Adjusted for calendar year, age, BMI and smoking	Unadjusted	Adjusted for calendar year	Adjusted for calendar year, age, BMI and smoking
1 y	136	1.1 (-0.2;2.4)	1.2 (-0.2;2.6)	0.8 (-2.4;4.0)	4.0 (-1.1;9.1)	3.5 (-1.7;8.7)
3 y	281	3.0 (1.0;5.0)	2.0 (-0.2;4.2)	4.1 (-0.8;9.0)	3.8 (-3.7,11.4)	3.0 (-4.7;10.6)

5

Cox proportional hazards and IV for Cox proportional hazards regression

Median follow-up time was 234 days, 38% had at least 1 year and 11% had at least 3 years of continuous use of the same OC. There were 179 events during a continuous period of use of the same OC. Hazard ratios (HRs) for VTE of 3rd generation OCs compared with 2nd generation OCs obtained using conventional Cox proportional hazards regression and IV for Cox proportional hazards regression are displayed in Table 3. Both the conventional Cox regression (adjusted HR 1.62, 95% CI 1.16-2.27) and the IV Cox regression (fully adjusted HR 3.45, 95% CI 0.97-11.7) indicated an increased VTE risk for 3rd generation OCs.

Table 3. *Conventional and instrumental variable estimates of the hazard ratio of venous thromboembolism for 3rd generation versus 2nd generation oral contraceptives.*

Conventional Cox proportional hazards regression		IV for Cox proportional hazards regression*		
Unadjusted	Adjusted for calendar year, age, BMI and smoking	Unadjusted	Adjusted for calendar year	Adjusted for calendar year, age, BMI and smoking
1.78 (1.30-2.44)	1.62 (1.16-2.27)	2.05 (0.96-4.39)	4.16 (1.22-13.0)	3.45 (0.97-11.7)

*IV regression with a linear regression first stage and a Cox regression second stage, confidence intervals derived using bootstrapping (see Methods for details)

Sensitivity analyses
The sensitivity analysis requiring a minimal registration time of 1 year gave slightly larger 1-year risk difference estimates in both the OLS analysis (adjusted RD 1.7, 95% CI 0.2;3.3) and the 2-SLS analysis (fully adjusted RD 5.7, 95% CI 0.0;11.4). Other results did not change materially (Supplementary Tables 2-4).

The sensitivity analysis without requirement of a record of anticoagulant treatment in the event definition resulted in a larger absolute number of events, and somewhat larger risk differences in the OLS analyses (Supplementary Table 5). All other results (Supplementary Tables 5 and 6) were very similar to those from the main analyses.

Discussion

The results of the IV analyses showed a similar picture to results from the conventional analyses. All estimates were consistently in the direction of an increased risk of VTE for 3rd generation OCs in comparison to 2nd generation OCs. The point estimates from the IV analysis were generally higher than those from the conventional analyses, albeit with wide confidence intervals. The results of the IV analyses do not indicate that unknown confounding could explain the higher VTE incidence with 3rd generation OCs.

To our knowledge, no previous studies have used IV analysis to investigate the effect of 3rd vs 2nd generation OCs on the risk of VTE. Previous studies include both case-control studies,[7;8;14;17-19] and cohort studies.[9;14;17;20] Many of these studies compared levonorgestrel-containing contraceptives with gestodene- or desogestrel-containing contraceptives,[7;9;17;20] whereas we included a broader range of 2nd and 3rd generation OCs (although there has been discussion whether norgestimate-containing OCs should be grouped with 3rd generation OCs)[12]. To mimic the randomised trial situation, we only used 'incident users' of OCs in our analysis.[21-23] For the Cox proportional hazards regression analyses we only included the period of use of the class of OC a patient was first prescribed. This highlights a limitation of the least squares analyses: these included women who started using a certain OC, but switched, stopped or were lost to follow-up.

A limitation of IV analysis in general is the larger variance in comparison to conventional analyses. Although our study population was very large, the number of events was small, resulting in large confidence intervals for the IV estimates in particular. A further limitation is the difficulty of identifying all true VTE events. We used an extensive

list of diagnosis codes, and required a record of an anticoagulant prescription for the events in our main analyses as an additional safeguard against misclassification (as in previous studies). However, this may have resulted in exclusion of some true events. Sensitivity analyses without the anticoagulant requirement did not yield substantially different results.

The analysis across the twenty-five year time period and the changes in prescribing preferences over time posed some problems. Because both prescribing preference and the age and smoking behaviour of patients who were first prescribed an OC changed substantially over time, prescribing preference was related to age and smoking behaviour. This violates the independence assumption (the instrument may not be related to other factors which affect the outcome). Restrict the data to a shorter time period across which prescribing preference was more or less stable was not possible due to the low incidence of VTE which would result in a study with a very low power. Adjusting the IV analyses for calendar year was considered the best alternative.

The similarity of the results from the conventional and IV analysis suggests major confounding is unlikely. If anything, the larger point estimates from the IV analyses suggest the conventional estimates are conservative, and do not support the objection by some researchers in the late 1990s that the higher VTE incidence for 3rd generation contraceptives was due to selective prescribing of 3rd generation OCs to women at an increased risk of VTE. Selective prescribing of 3rd generation OCs to women at *low* risk of VTE after 1995 could have resulted in some degree of confounding by contra-indication. In general, the degree to which confounding by contra-indication is likely to be present will depend on the predictability of the adverse effect studied. For example, whereas the risk of bleeding upon use of oral anti-coagulation is relatively predictable, the risk of developing a cough upon use of angiotensin-converting enzyme inhibitors is essentially unpredictable.

In conclusion, we found an increased risk of VTE for 3rd generation OCs in comparison to 2nd generation OCs using both IV analyses and conventional analyses. We conclude that major confounding is unlikely in this study of a minimally predictable side-effect. IV analysis can be a useful sensitivity analysis to assess the presence of confounding in studies of adverse effects in very large databases.

References

(1) Greenland S, Neutra R. Control of confounding in the assessment of medical technology. *Int J Epidemiol* 1980;9:361-367.

(2) Vandenbroucke JP. When are observational studies as credible as randomised trials? *Lancet* 2004;363:1728-1731.

(3) Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. *PLoS Med* 2011;8:e1001026.

(4) Papanikolaou PN, Christidi GD, Ioannidis JP. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. *CMAJ* 2006;174:635-641.

(5) Feenstra H, Grobbee RE, in't Veld BA, Stricker BH. Confounding by contraindication in a nationwide cohort study of risk for death in patients taking ibopamine. *Ann Intern Med* 2001;134:569-572.

(6) Miettinen OS. The need for randomization in the study of intended effects. *Stat Med* 1983;2:267-271.

(7) Effect of different progestagens in low oestrogen oral contraceptives on venous thromboembolic disease. World Health Organization Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception. *Lancet* 1995;346:1582-1588.

(8) Bloemenkamp KW, Rosendaal FR, Helmerhorst FM, Buller HR, Vandenbroucke JP. Enhancement by factor V Leiden mutation of risk of deep-vein thrombosis associated with oral contraceptives containing a third-generation progestagen. *Lancet* 1995;346:1593-1596.

(9) Jick H, Jick SS, Gurewich V, Myers MW, Vasilakis C. Risk of idiopathic cardiovascular death and nonfatal venous thromboembolism in women using oral contraceptives with differing progestagen components. *Lancet* 1995;346:1589-1593.

(10) Vandenbroucke JP, Rosing J, Bloemenkamp KW et al. Oral contraceptives and the risk of venous thrombosis. *N Engl J Med* 2001;344:1527-1535.

(11) Kemmeren JM, Algra A, Grobbee DE. Third generation oral contraceptives and risk of venous thrombosis: meta-analysis. *BMJ* 2001;323:131-134.

(12) Walker AM. Newer oral contraceptives and the risk of venous thromboembolism. *Contraception* 1998;57:169-181.

(13) Farley TM, Meirik O, Collins J. Cardiovascular disease and combined oral contraceptives: reviewing the evidence and balancing the risks. *Hum Reprod Update* 1999;5:721-735.

(14) Farmer RD, Lawrenson RA, Thompson CR, Kennedy JG, Hambleton IR. Population-based study of risk of venous thromboembolism associated with various oral contraceptives. *Lancet* 1997;349:83-88.

(15) Small DS, Tan Z, Lorch SA, Brookhart MA. Instrumental variable estimation when compliance is not deterministic: the stochastic monotonicity assumption. 2014.

(16) Tchetgen Tchetgen EJ, Walter S, Vansteelandt S, Martinussen T, Glymour M. Instrumental Variable Estimation in a Survival Context. *Epidemiology* 2015.

(17) Jick H, Kaye JA, Vasilakis-Scaramozza C, Jick SS. Risk of venous thromboembolism among users of third generation oral contraceptives compared with users of oral contraceptives with levonorgestrel before and after 1995: cohort and case-control analysis. *BMJ* 2000;321:1190-1195.
(18) Spitzer WO, Lewis MA, Heinemann LA, Thorogood M, MacRae KD. Third generation oral contraceptives and risk of venous thromboembolic disorders: an international case-control study. Transnational Research Group on Oral Contraceptives and the Health of Young Women. *BMJ* 1996;312:83-88.
(19) Lidegaard O, Edstrom B, Kreiner S. Oral contraceptives and venous thromboembolism. A case-control study. *Contraception* 1998;57:291-301.
(20) Herings RM, Urquhart J, Leufkens HG. Venous thromboembolism among new users of different oral contraceptives. *Lancet* 1999;354:127-128.
(21) Hernan MA. The hazards of hazard ratios. *Epidemiology* 2010;21:13-15.
(22) Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol* 2003;158:915-920.
(23) Danaei G, Tavakkoli M, Hernan MA. Bias in observational studies of prevalent users: lessons for comparative effectiveness research from a meta-analysis of statins. *Am J Epidemiol* 2012;175:250-262.

Supplementary Table 1. *First-stage associations of three potential instruments with actual prescription of a 3rd generation oral contraceptive.*

Potential instrument	Difference in probability of prescription*	Partial r^2†	Partial F-statistic†
Previous prescription			
Unadjusted	0.308	0.095	44626
Adjusted for calendar year	0.191	0.037	16173
Adjusted for calendar year, age, BMI and smoking	0.190	0.036	15859
Proportion of past 5 prescriptions for 3rd generation oral contraceptives			
Unadjusted	0.628	0.184	96055
Adjusted for calendar year	0.464	0.081	37369
Adjusted for calendar year, age, BMI and smoking	0.461	0.080	36728
Proportion of prescriptions in past year for 3rd generation oral contraceptives			
Unadjusted	0.669	0.191	99357
Adjusted for calendar year	0.503	0.085	38881
Adjusted for calendar year, age, BMI and smoking	0.498	0.084	38293

*Difference in probability of a prescription for a 3rd generation oral contraceptive for patients with a GP whose previous prescription was a 3rd generation oral contraceptive versus patients with a GP whose previous prescription was a 2nd generation oral contraceptive. Difference in probability of a prescription for a 3rd generation oral contraceptive for patients with a GP whose previous 5 prescriptions were all 3rd generation oral contraceptives versus patients with a GP whose previous 5 prescriptions were all 2nd generation oral contraceptives. Difference in risk of a prescription for a 3rd generation oral contraceptive for patients with a GP whose prescriptions in the past year were all 3rd generation oral contraceptives versus patients with a GP whose prescriptions in the past year were all 2nd generation oral contraceptives.

†For the fully adjusted analysis the partial r2 and partial F-statistic were approximated by their averages across 10 imputations.

Supplementary Table 2. *Patient characteristics by actual type of oral contraceptive and by quintiles of prescriptions for 3rd generation oral contraceptives by their GP in the past year for patients with a minimal time between entry date and date of first prescription of 1 year.*

	Actual prescription		Prescriptions of the same GP in past year (% 3rd generation)			
	2nd generation	3rd generation	0	Q1 (1.5-21.9)	Q2 (21.9-50.0)	Q3 (50.5-100)
N	265366	90760	133200	74342	86862	61722
Actual prescription 3rd generation	N/A	N/A	14571 (10.9)	11054 (14.9)	23848 (27.5)	41287 (66.9)
Age (y), median (IQR)	19.3 (16.8-27.5)	23.6 (18.0-30.6)	19.1 (16.8-27.0)	19.9 (17.0-28.2)	21.0 (17.2-29.2)	22.7 (17.7-29.9)
BMI (kg/m^2), median (IQR)†	22.8 (20.5-26.2)	22.8 (20.6-25.8)	22.9 (20.5-26.2)	22.9 (20.6-26.3)	22.8 (20.5-25.9)	22.6 (20.5-25.6)
Smoking‡						
Yes	51588 (24.4)	17831 (26.2)	25386 (23.3)	14937 (25.3)	17158 (25.4)	11938 (27.5)
Ex	15324 (7.3)	5661 (8.3)	8323 (7.6)	4489 (7.6)	5255 (7.8)	2918 (6.7)
Venous thrombo-embolism						
1 year	77 (0.03)	40 (0.05)	43 (0.03)	22 (0.03)	26 (0.03)	26 (0.04)
3 years	155 (0.07)	85 (0.10)	80 (0.08)	52 (0.08)	52 (0.07)	56 (0.10)

Data are presented as n(%) unless stated otherwise.
†Data available for 192252 patients.
‡Data available for 279155 patients.

Supplementary Table 3. *Conventional and instrumental variable estimates of the risk differences of venous thromboembolism per 10000 patients for 3rd generation versus 2nd generation oral contraceptives within 1 year and within 3 years of first prescription, for patients with a minimal time between entry date and date of first prescription of 1 year.*

		Ordinary least squares (Risk difference per 10,000)		Two-stage least squares (Risk difference per 10,000)		
Time	No. Events	Unadjusted	Adjusted for calendar year, age, BMI and smoking	Unadjusted	Adjusted for calendar year	Adjusted for calendar year, age, BMI and smoking
1 year	117	1.5 (0.0;2.9)	1.7 (0.2;3.3)	2.0 (-1.4;5.5)	6.2 (0.5;11.9)	5.7 (0.0;11.4)
3 year	240	3.2 (1.0;5.4)	2.2 (-0.2;4.7)	4.4 (-0.8;9.7)	4.7 (-3.3;12.8)	3.7 (-4.4;11.9)

Supplementary Table 4. *Conventional and instrumental variable estimates of the hazard ratio of venous thromboembolism for 3rd generation versus 2nd generation oral contraceptives, for patients with a minimal time between entry date and date of first prescription of 1 year.*

Conventional Cox proportional hazards regression		IV for Cox proportional hazards regression*		
Unadjusted	Adjusted for calendar year, age, BMI and smoking	Unadjusted	Adjusted for calendar year	Adjusted for calendar year, age, BMI and smoking
1.86 (1.31-2.62)	1.69 (1.17;2.44)	2.25 (0.96-4.93)	4.57 (1.33-14.45)	3.77 (1.07;11.8)

*IV regression with a linear regression first stage and a Cox regression second stage, confidence intervals derived using bootstrapping (see Methods for details).

Supplementary Table 5. *Conventional and instrumental variable estimates of the risk differences of venous thromboembolism per 10000 patients for 3rd generation versus 2nd generation oral contraceptives within 1 year and within 3 years of first prescription, without requirement of a record of anticoagulant prescription for the thromboembolic events.*

		Ordinary least squares (Risk difference per 10,000)		Two-stage least squares (Risk difference per 10,000)		
Time	No. Events	Unadjusted	Adjusted for calendar year, age, BMI and smoking	Unadjusted	Adjusted for calendar year	Adjusted for calendar year, age, BMI and smoking
1 year	209	2.3 (0.7;3.9)	2.0 (0.2;3.7)	2.7 (-1.4;6.7)	4.0 (-1.7;9.8)	3.5 (-2.4;9.3)
3 year	453	4.2 (1.7;6.8)	2.4 (-0.4;5.1)	6.8 (0.6;13.0)	3.3 (-5.8;12.4)	2.4 (-6.8;11.6)

Supplementary Table 6. *Conventional and instrumental variable estimates of the hazard ratio of venous thromboembolism for 3rd generation versus 2nd generation oral contraceptives, without requirement of a record of anticoagulant prescription for the thromboembolic events.*

Conventional Cox proportional hazards regression		IV for Cox proportional hazards regression*		
Unadjusted	Adjusted for calendar year, age, BMI and smoking	Unadjusted	Adjusted for calendar year	Adjusted for calendar year, age, BMI and smoking
1.74 (1.34;2.24)	1.59 (1.22;2.09)	2.11 (1.19;3.75)	3.53 (1.42;8.66)	3.03 (1.16;7.19)

*IV regression with a linear regression first stage and a Cox regression second stage, confidence intervals derived using bootstrapping (see Methods for details).

Supplementary Table 7. *Patient characteristics by calendar year.*

	Year				
	1989-1993	1994-1998	1999-2003	2004-2008	2009-2013
N	35861	50760	107205	126810	99516
Age (y), median (IQR)	24.7 (19.5-29.8)	23.7 (18.0-30.1)	22.3 (17.6-29.8)	20.4 (17.0-28.3)	19.4 (16.9-26.9)
BMI (kg/m^2), median (IQR)†	22.5 (20.6-25.3)	22.5 (20.5-25.4)	22.9 (20.7-26.1)	23.1 (20.7-26.4)	23.0 (20.5-26.4)
Smoking, n(%)‡					
Yes	6797 (32.1)	10859 (30.0)	21936 (30.4)	28344 (25.3)	17191 (18.4)
Ex	1074 (5.1)	2224 (6.1)	6100 (8.5)	10845 (9.7)	7861 (8.4)

†Data available for 239593 patients.
‡Data available for 335553 patients.

Supplementary Codelists

Codelist 1. *British National Formulary codes for combined hormonal contraceptives*

Product code	Product name	BNF code
	Oral combined hormonal contraceptives	
	2nd generation	
41	Microgynon 30 tablets (Bayer Plc)	7030152
978	Logynon tablets (Bayer Plc)	7030152
1062	Ovranette 150microgram/30microgram tablets (Pfizer Ltd)	7030152
2026	Logynon ED tablets (Bayer Plc)	7030152
2084	Ovran 30 Tablet (Wyeth Pharmaceuticals)	7030100
4917	Microgynon 30 ED tablets (Bayer Plc)	7030152
6686	Ethinylestradiol 30microgram / Levonorgestrel 150microgram tablets	7030152
12631	Ethinylestradiol with levonorgestrel and placebo 30micrograms + 150micrograms Tablet	7030100
15886	Ethinylestradiol with levonorgestrel - triphasic 6x30+50mcg; 5x40+75mcg; 10x30+125mcg Tablet	7030100
23897	Ethinylestradiol with levonorgestrel - triphasic with placebo 6x30+50mcg; 5x40+75mcg; 10x30+125mcg Tablet	7030100
42510	Ethinylestradiol with levonorgestrel Tablet	7030100
43003	Levest 150/30 tablets (Morningside Healthcare Ltd)	7030152
44046	Rigevidon tablets (Consilient Health Ltd)	7030152
44278	TriRegol tablets (Consilient Health Ltd)	7030152
45059	Ethinylestradiol with levonorgestrel 30micrograms + 50micrograms Tablet	7030100
45557	Levest 150/30 tablets (Actavis UK Ltd)	7030152
47281	Elevin 150microgram/30microgram tablets (MedRx Healthcare LLP)	7030152
51482	Generic Microgynon 30 ED tablets	7030152
52443	Microgynon 30 tablets (Mawdsley-Brooks & Company Ltd)	7030152
56868	Generic Logynon ED tablets	7030152
1352	Loestrin 30 tablets (Galen Ltd)	7030150
1354	Brevinor 500microgram/35microgram tablets (Pfizer Ltd)	7030152
1427	Loestrin 20 tablets (Galen Ltd)	7030150
1601	Trinovum tablets (Janssen-Cilag Ltd)	7030100
1988	Binovum tablets (Janssen-Cilag Ltd)	7030152
2354	Ovysmen 500microgram/35microgram tablets (Janssen-Cilag Ltd)	7030152
2856	Norimin 1mg/35microgram tablets (Pfizer Ltd)	7030152
3472	Trinovum ed ED tablets (Janssen-Cilag Ltd)	7030100
3538	Neocon 1/35 Tablet (Cilag Pharmaceuticals Ltd)	7030100
5576	Synphase tablets (Pfizer Ltd)	7030152

7814	Ethinylestradiol 35microgram / Norethisterone 500microgram tablets	7030152
8176	Ethinylestradiol 20microgram / Norethisterone acetate 1mg tablets	7030150
8482	Ethinylestradiol 35microgram / Norethisterone 1mg tablets	7030152
14670	Ethinylestradiol with norethisterone - biphasic 7 x 35mcg+500mcg; 14 x 35mcg+1mg Tablet	7030100
15987	Ethinylestradiol with norethisterone - triphasic 7 x 35+500mcg; 7 x 35+750mcg; 7 x 35mcg+1mg Tablet	7030100
18823	Ethinylestradiol 30microgram / Norethisterone acetate 1.5mg tablets	7030150
31528	Ethinylestradiol with norethisterone - triphasic 7x35+500mcg; 9x35mcg+1mg; 5x35+500mcg Tablet	7030100
40650	Ethinylestradiol with norethisterone 35micrograms + 750micrograms Tablet	7030100
56553	Generic Trinovum tablets	7030100
	3rd generation	
443	Desogestrel with ethinylestradiol 150micrograms with 20micrograms tablets	7030100
935	Marvelon tablets (Merck Sharp & Dohme Ltd)	7030152
1378	Mercilon 150microgram/20microgram tablets (Merck Sharp & Dohme Ltd)	7030150
13248	Ethinylestradiol 30microgram / Desogestrel 150microgram tablets	7030152
16110	Ethinylestradiol 20microgram / Desogestrel 150microgram tablets	7030150
23211	Desogestrel with ethinylestradiol 150micrograms with 30micrograms tablets	7030100
44336	Gedarel 30microgram/150microgram tablets (Consilient Health Ltd)	7030152
44457	Gedarel 20microgram/150microgram tablets (Consilient Health Ltd)	7030150
49700	Marvelon tablets (Doncaster Pharmaceuticals Ltd)	7030152
58642	Cimizt 30microgram/150microgram tablets (Morningside Healthcare Ltd)	7030152
936	Femodene tablets (Bayer Plc)	7030150
977	Minulet tablets (Wyeth Pharmaceuticals)	7030150
3471	Femodene ED tablets (Bayer Plc)	7030152
3693	Triadene tablets (Bayer Plc)	7030152
4964	Femodette tablets (Bayer Plc)	7030150
7776	Ethinylestradiol 30microgram / Gestodene 75microgram tablets	7030150
11910	Ethinylestradiol 20microgram / Gestodene 75microgram tablets	7030150
14977	Ethinylestradiol with gestodene - triphasic 6 x 30+50mcg; 5 x 40+70mcg; 10 x 30+100mcg Tablet	7030100
18569	Gestodene with ethinylestradiol 75microgramwith20microgram Tablet	7030100

19131	Ethinylestradiol with gestodene and placebo 30micrograms + 75micrograms Tablet	7030100
21733	Gestodene with ethinylestradiol 75microgramwith30microgram Tablet	7030100
36829	Katya 30/75 tablets (Stragen UK Ltd)	7030150
37073	Sunya 20/75 tablets (Stragen UK Ltd)	7030150
44229	Millinette 20microgram/75microgram tablets (Consilient Health Ltd)	7030150
44994	Millinette 30microgram/75microgram tablets (Consilient Health Ltd)	7030150
56483	Generic Tri-Minulet tablets	7030152
1071	Cilest 250microgram/35microgram tablets (Janssen-Cilag Ltd)	7030152
14601	Ethinylestradiol 35microgram / Norgestimate 250microgram tablets	7030152
25263	Norgestimate with ethinylestradiol 250micrograms + 35micrograms Tablet	7030100
49214	Cilest 250microgram/35microgram tablets (Mawdsley-Brooks & Company Ltd)	7030152
57181	Lizinna 250microgram/35microgram tablets (Morningside Healthcare Ltd)	7030152
	drospirenone-containing	
697	Yasmin tablets (Bayer Plc)	7030152
6716	Ethinylestradiol 30microgram / Drospirenone 3mg tablets	7030152
47057	Yasminelle 3mg+20microgram Tablet (Bayer Plc)	7030100
52818	Yasmin tablets (Mawdsley-Brooks & Company Ltd)	7030152
56311	Yaz tablets (Imported (United States))	07030152/ 13060202
	Other	
125	Dianette tablets (Bayer Plc)	13090200/ 13060202/ 07030100
2769	Cyproterone acetate with ethinylestradiol 2mg with 35micrograms tablets	07030100/ 13060202
4608	Dianette tablets (Generics (UK) Ltd)	13060202/ 13090200/ 07030100
6431	Co-cyprindiol 2000microgram/35microgram tablets	07030100/ 13060202/ 13090200
22603	Clairette 2000/35 tablets (Stragen UK Ltd)	07030100/ 13060202/ 13090200
23218	Ethinylestradiol with cyproterone acetate 35microgram with 2mg tablets	07030100/ 13060202

25124	Acnocin 2000microgram/35microgram tablets (Sandoz Ltd)	13060202/ 13090200/ 07030100
31902	Cicafem 2000/35 tablets (Galen Ltd)	13090200/ 07030100/ 13060202
33098	Diva 2000/35 tablets (Zeroderma Ltd)	07030100/ 13060202/ 13090200
38500	Co-cyprindiol 2000microgram/35microgram tablets (Fannin UK Ltd)	13090200/ 07030100/ 13060202
47132	Co-cyprindiol 2mg+35microgram Tablet (Sandoz Ltd)	13060202/ 07030100
53201	Dianette tablets (Lexon (UK) Ltd)	07030100/ 13090200/ 13060202
8103	Conova 30 Tablet (Pharmacia Ltd)	7030100
40305	Qlaira tablets (Bayer Plc)	7030100
40618	Estradiol valerate and (estradiol valerate with dienogest) tablets	7030100
56539	Zoely 2.5mg/1.5mg tablets (Merck Sharp & Dohme Ltd)	7030152
57264	Estradiol 1.5mg / Nomegestrol 2.5mg tablets	7030152
3436	Ortho-novin 1/50 Tablet (Janssen-Cilag Ltd)	7030100
5862	Norinyl-1 tablets (Pfizer Ltd)	7030152
9119	Minilyn Tablet (Organon Laboratories Ltd)	7030100
14459	Gynovlar 21 Tablet (Schering Health Care Ltd)	7030100
17756	Mestranol 50microgram / Norethisterone 1mg tablets	7030152
19551	Controvlar Tablet (Schering Health Care Ltd)	7030100
21343	Minovlar ed Tablet (Schering Health Care Ltd)	7030100
43009	Minovlar 21 Tablet (Schering Health Care Ltd)	7030100
	Vaginal combined hormonal contraceptives	
39517	NuvaRing 0.12mg/0.015mg per day vaginal delivery system (Merck Sharp & Dohme Ltd)	07030151
40512	Ethinylestradiol 2.7mg / Etonogestrel 11.7mg vaginal delivery system	07030151
	Transdermal patches	
6166	Evra transdermal patches (Janssen-Cilag Ltd)	07030101
6596	Norelgestromin with ethinylestradiol 203micrograms + 33.9micrograms/24hours Transdermal patch	07030100
29499	Ethinylestradiol 33.9micrograms/24hours / Norelgestromin 203micrograms/24hours transdermal patches	07030101

Codelist 2. *Oxford Medical Information Systems and Read codes used to define the outcome of venous thromboembolism.*

Read Code	Description
	Deep vein thrombosis
G801.11	Deep vein thrombosis
G801.12	Deep vein thrombosis, leg
G801.13	DVT- Deep vein thrombosis
G801C00	Deep vein thrombosis of leg related to air travel
G801D00	Deep vein thrombosis of lower limb
G801E00	Deep vein thrombosis of leg related to intravenous drug use
G801F00	Deep vein thrombosis of peroneal vein
G822.00	Embolism and thrombosis of the vena cava
G824.00	Axillary vein thrombosis
G825.00	Thrombosis of subclavian vein
	Thrombophlebitis
G801.00	Deep vein phlebitis and thrombophlebitis of the leg
G801500	Deep vein phlebitis of the leg unspecified
G801600	Thrombophlebitis of the femoral vein
G801700	Thrombophlebitis of the popliteal vein
G801800	Thrombophlebitis of the anterior tibial vein
G801A00	Thrombophlebitis of the posterior tibial vein
G801B00	Deep vein thrombophlebitis of the leg unspecified
G801z00	Deep vein phlebitis and thrombophlebitis of the leg NOS
G80y.11	Phlebitis and/or thrombophlebitis of iliac vein
G80y400	Thrombophlebitis of the common iliac vein
G80y500	Thrombophlebitis of the internal iliac vein
G80y600	Thrombophlebitis of the external iliac vein
G80y700	Thrombophlebitis of the iliac vein unspecified
G80y800	Phlebitis and thrombophlebitis of the iliac vein NOS
	Pulmonary embolism
G401.00	Pulmonary embolism
G401.12	Pulmonary embolus
G401000	Post operative pulmonary embolus

Codelist 3. *Oxford Medical Information Systems and Read codes used to define a history of venous thromboembolism (deep vein thrombosis or pulmonary embolism), in addition to the codes listed in Table 2.*

Read Code	Description
14A8100	H/O: Deep Vein Thrombosis
14A8.12	H/O: Thrombosis
ZV12900	[V] Personal history of pulmonary embolism
ZV12811	[V] Personal history DVT- deep vein thrombosis
14A8.00	H/O: thrombo-embolism
14A8.11	H/O: embolism
ZV12800	[V] Personal history deep vein thrombosis
14AC.00	H/O: pulmonary embolus

Chapter 6

Reporting instrumental variable analyses.

Anna G.C. Boef, Olaf M. Dekkers, Saskia le Cessie and Jan P. Vandenbroucke

Epidemiology. 2013 Nov;24(6):937-8

To the editor:

Swanson and Hernan[1] discuss the reporting of instrumental variable analyses, and propose a step-by-step checklist. Two phases can be distinguished in their suggested steps. The first comprises discussion of the three main instrumental variable assumptions. The second phase concerns estimation of the effect; they propose that authors should first discuss whether the effect in the population or the effect in the compliers is of interest, should then estimate bounds for the effect and should finally, if appropriate, justify an additional assumption that allows estimation of a point estimate.

We would like to suggest an intermediate reporting step, in between these two phases: to present the distribution of the outcome across instrumental variable values. This amounts to a crude analysis of the effect of levels of the instrumental variable on the outcome. It resembles the form of the usual epidemiologic study in which two or more groups are contrasted, with different levels of exposure frequency.

For a dichotomous instrument the presentation of the outcome across values of the instrument is straightforward. The comparison of the outcome between the two values of the instrument gives an effect estimate which can be thought of as similar to an intention-to-treat effect in a randomised trial. Davies et al provide a specific reporting suggestion in case instrument, treatment and outcome are all dichotomous, namely tabulation of frequencies of all combinations of instrument treatment and outcome.[2] An example of how our suggested step can be reported if the instrument is continuous, is presented in a paper by Stukel et al.[3] They performed an instrumental variable analysis with regional cardiac catheterisation rates as an instrument to investigate the effect of cardiac catheterisation (as a marker of intent to treat invasively) on long-term survival in acute myocardial infarction. They provide a table [Table 4 in their paper] with baseline characteristics as well as the outcome (mortality) across quintiles of regional cardiac catheterisation rate. Direct comparison of the outcome across instrument quintiles shows a decrease in mortality with increasing regional cardiac catheterisation rate. The display of baseline characteristics across the same quintiles allows the reader to evaluate how comparable patient characteristics in these quintiles are (third instrumental variable assumption: the instrument is independent of confounders[2]).

Such an additional step in the reporting of instrumental variable analyses provides a presentation of the data before a decision is made about reporting bounds only or a point estimate. An intermediate analysis might be done on these data, e.g. by doing a comparative analysis of a dichotomous instrumental variable, or by directly contrasting

the lowest and highest categories of a continuous instrumental variable distribution. The validity of this comparison does of course depend on the three main instrumental variable assumptions and violations of these assumptions will lead to bias. However, the bias *amplification* which can occur when using standard instrumental variable methods to obtain effect estimates[4] will not affect the comparison of the outcome across strata of the instrument. Falsification tests of the third instrumental variable assumption are discussed by Swanson and Hernan[1] and Davies et al[2]. Showing the distribution of patient characteristics across values of the instrument (in parallel to the distribution of the outcome across values of the instrument) may also aid in detecting potential violations of this assumption.

References

(1) Swanson SA, Hernan MA. Commentary: how to report instrumental variable analyses (suggestions welcome). *Epidemiology* 2013;24:370-374.
(2) Davies NM, Smith GD, Windmeijer F, Martin RM. Issues in the reporting and conduct of instrumental variable studies: a systematic review. *Epidemiology* 2013;24:363-369.
(3) Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA* 2007;297:278-285.
(4) Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006;17:360-372.

Chapter 7

Mendelian randomization studies: a review of the approaches used and the quality of reporting.

Anna G.C. Boef, Olaf M. Dekkers and Saskia le Cessie

Int J Epidemiol. 2015 Apr; 44(2):496-511

Abstract

Background: Mendelian randomisation (MR) studies investigate the effect of genetic variation in levels of an exposure on an outcome, thereby using genetic variation as an instrumental variable (IV). We provide a meta-epidemiological overview of the methodological approaches used in MR studies, and evaluate the discussion of MR assumptions and reporting of statistical methods.

Methods: We searched PubMed, Medline, Embase and Web of Science for MR studies up to December 2013. We assessed 1) the MR approach used; 2) whether the plausibility of MR assumptions was discussed, and 3) whether the statistical methods used were reported adequately.

Results: Of 99 studies using data from one study population, 32 used genetic information as a proxy for the exposure without further estimation, 44 performed a formal IV analysis, 7 compared the observed with the expected genotype-outcome association, and 1 used both the latter two approaches. The 80 studies using data from multiple study populations used many different approaches to combine the data. Fifty-two of these studies used some form of IV analysis. Forty-four percent of studies discussed the plausibility of all three MR assumptions in their study. Statistical methods used for IV analysis were insufficiently described in 14% of studies.

Conclusions: Most MR studies either use the genotype as a proxy for exposure without further estimation or perform an IV analysis. The discussion of underlying assumptions and reporting of statistical methods for IV analysis are frequently insufficient. Studies using data from multiple study populations are further complicated by the combination of data or estimates. We provide a checklist for the reporting of MR studies.

Key words: Mendelian randomisation, instrumental variable, aetiology.

Medical Subject Headings: Mendelian Randomization Analysis; Genetic Variation; Confounding Factors (Epidemiology); Causality.

7

Key messages:

- The specific methods used in Mendelian randomisation studies vary widely.
- These methods broadly fall into three categories: 1) using genetic information as a proxy for the exposure without further estimation, 2) performing an instrumental variable analysis; 3) comparing the observed with the expected genotype-outcome association.
- Mendelian randomisation studies frequently insufficiently discuss underlying assumptions and report statistical methods for IV analysis.
- A checklist for the reporting of Mendelian randomisation studies is provided.

Introduction

Observational studies are limited in their ability to identify whether exposures are causally related to disease occurrence or other outcomes. Adjustment for confounding is only possible for those factors which are identified and measured and will inevitably be incomplete: some degree of residual confounding will always remain. Reverse causation, an effect of the outcome on the studied exposure, may also explain associations found in an observational study.[1,2] An approach which can circumvent both reverse causation (as first proposed in 1986)[3] and residual confounding in order to establish the causal effect of the exposure on the outcome is to investigate the effect of genetic variation in levels of the exposure on the outcome. This approach has come to be known as Mendelian randomisation over the last decade.[2] The random allocation of genetic variants from parents to offspring means these variants will generally be unrelated to other factors which affect the outcome.[1] Furthermore, associations between the genotype and the outcome will not be affected by reverse causation because disease does not affect genotype.[1]

Mendelian randomisation studies use genetic variation as an instrumental variable (IV) and must fulfil instrumental variable assumptions. Applied to Mendelian randomisation, these assumptions are that (1) the genotype is associated with the exposure; (2) the genotype is associated with the outcome through the studied exposure only (exclusion restriction assumption); and (3) the genotype is independent of other factors which affect the outcome (independence assumption).[4] Potential threats to the validity of these assumptions, such as population stratification, linkage disequilibrium, and pleiotropic effects are discussed in detail elsewhere.[1,5]

These general principles of Mendelian randomisation are increasingly used in aetiologic research, but the specific methods used in these studies can vary widely. In this study we review the methodology used in studies from the past 10 years which were described by the authors as Mendelian randomisation studies. We provide an overview of the use of the different approaches to Mendelian randomisation and where applicable the specific statistical methods used for estimation. We evaluate whether the plausibility of the Mendelian randomisation assumptions is discussed. Further we evaluate whether the statistical methods used are sufficiently described (including how the confidence interval was obtained) for those studies which perform an instrumental variable analysis or compare the observed and expected genotype-outcome association.

Methods

Search strategy and inclusion criteria

We searched PubMed, Medline, Embase and Web of Science for studies containing the term "Mendelian randomisation" or "genetic instrumental variable" or a related term (e.g. "genetic instrument") from January 1st 2003 to December 31st 2013. The full search strategies for each of the databases are included in the Supplementary Methods. We excluded publications that (1) were conference abstracts, letters, commentaries, editorials, reviews, study proposals or theoretical papers; (2) did not use Mendelian randomisation (i.e. did not state Mendelian randomisation or a genetic instrumental variable was used in the text, abstract or title and did not include "Mendelian randomisation" or "genetic instrumental variable" or a related term as a keyword); (3) identified potential genetic instruments for future Mendelian randomisation studies; (4) were primarily methodological using an application of Mendelian randomisation as an example; or (5) were published in a health economics journal (rather than a biomedical journal).

Classification of Mendelian randomisation approach used

First we classified publications into studies which used data from a single study population and studies which used data from multiple study populations. We then classified included studies according to their general Mendelian randomisation approach: i.e. how they utilised the genetically determined variation in exposure.

A. For studies performed in a single study population we identified the following three main approaches:
   1. Use of genetic variation as a proxy for the exposure, without further estimation. These studies investigate the association between a genotype (which affects the exposure) and the outcome. No comparison is made to the *expected* association between this genotype and the outcome, and no IV estimate of the effect of the exposure on the outcome is obtained.
   2. Comparison of the observed and expected genotype-outcome associations. These studies compare the observed association between the genotype and the outcome to the association which would be expected if the observed exposure-outcome association were causal. This expected association is calculated by multiplying the observed genotype-exposure association with the observed exposure-outcome association (sometimes termed the 'triangulation' approach, although this is not a specific term). See Figure 1 for an illustration and further explanation. The confidence interval of the expected genotype-outcome association can be estimated analytically or using bootstrap techniques.

3. Formal instrumental variable analysis using genetic variation as the instrument. These studies perform a formal IV analysis to obtain a causal estimate of the effect of genetically determined variation in the exposure on the outcome. Different statistical techniques can be used for this purpose, as we will further explain below.

B. We classified the studies including more than one study population into the following pre-specified main categories:
   1. Pooling of the data, followed by any of the approaches A1-3 listed above.
   2. IV analysis in each of the study populations, followed by a meta-analysis.
   3. Meta-analysis using the genotype as a proxy for the exposure, without further estimation.
   4. Meta-analyses of the genotype-exposure, exposure-outcome and genotype-outcome associations, followed by comparison of observed and expected genotype-outcome associations (as in approach A2).

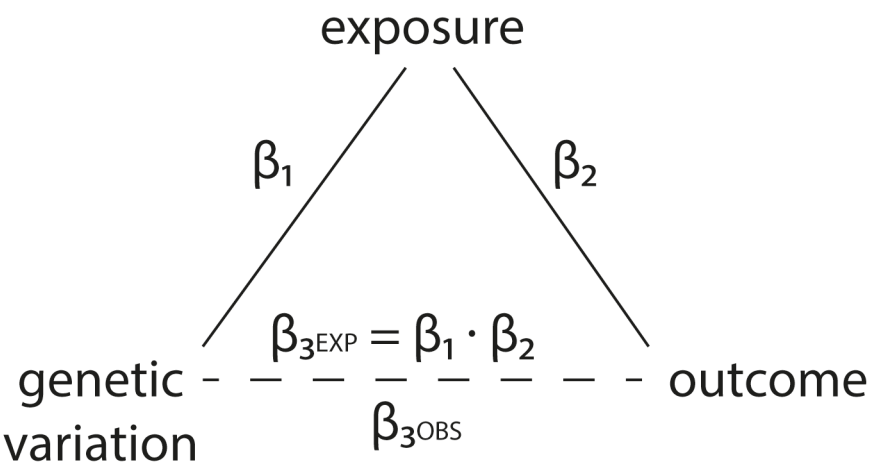


Figure 1. *Diagram of the approach used by Mendelian randomisation studies which compare the observed genotype-outcome association to the expected genotype-outcome association.*
β_1 regression coefficient of the genetic variant-exposure association.
β_2 regression coefficient of the exposure-outcome association.
β_{3OBS} observed regression coefficient of the genetic variant-outcome association.
β_{3EXP} expected regression coefficient of the genetic variant-outcome association.
The point estimate of β_{3EXP} is calculated as follows:
$\beta_{3EXP} = \beta_1 \cdot \beta_2$
The confidence interval of the expected genotype-outcome association can be estimated analytically or using bootstrap techniques.

5. Meta-analyses of the genotype-exposure and genotype-outcome associations, followed by a Wald-type/ratio estimate (see Didelez et al for a description of Wald-type estimators)[6] .
6. Data analysed separately for more than 1 population, followed by any of the approaches A1-3.

Further categories were added for those studies which did not fall into any of the above categories.

Assessment of discussion of Mendelian randomisation assumptions

Regardless of the approach used, Mendelian randomisation studies rely on three main assumptions as briefly mentioned in the introduction.

1. The genotype is associated with the exposure.
   This assumption can and should be verified in the data. Reporting guidelines for IV analyses recommend the use of the partial F-statistic as a measure of the strength of the association between the IV and the exposure.[7,8] It encompasses information on the strength of the instrument and on the number of observations in the analysis.[9] We assessed whether studies reported the strength of the genotype-exposure association in the data using a partial F-statistic or using another measure (e.g. mean difference in exposure by genotype). If not, we assessed whether they reported the strength of this association from literature.
2. The genotype is associated with the outcome through the studied exposure only (exclusion restriction assumption).
   This assumption is violated if the genotype has multiple (pleiotropic) effects, if a nearby variant with which it is in linkage disequilibrium affects the outcome in other ways than through the exposure of interest, or if developmental canalisation occurs.[1] For all studies we evaluated whether the plausibility of this assumption was discussed. Mentioning the assumption in general terms was not deemed sufficient: a specific discussion of its plausibility in the particular study was required.
3. The genotype is independent of other factors which affect the outcome (independence assumption).
   This assumption is violated if subgroups in the study population have both different genotype frequencies and different distributions of the outcome (population stratification).[1] It is also violated if there is an association between the genotype used as an instrument and confounders. For all studies we assessed whether the association between the genetic instrument and measured confounders was reported, as recommended in IV reporting guidelines.[7] Furthermore, we assessed whether potential associations of the genotype with unmeasured confounders were discussed and/or population stratification was discussed. Again, a specific discussion of the plausibility of the assumption in the particular study was required.

Assessment of reporting of statistical analysis.
This section only applies to the studies which used the IV approach or the observed-expected approach, because using genetic variation as a proxy for the exposure without further estimation does not involve any special statistical methods.

1. For studies which obtained an IV estimate of the effect of the exposure on the outcome we determined which statistical method was used, assessed whether it was described sufficiently, and whether a confidence interval was reported. A frequently used IV method is two-stage least squares analysis. This involves two stages of linear regression. The first stage is a linear regression with the exposure as the dependent variable and the instrument (genotype) as the independent variable, which is then used to obtain *predicted* exposure levels based on the instrument. The second stage is a regression with the outcome as the dependent variable and these genetically predicted exposure levels as the independent variable. Software for two-stage least squares regression takes into account the errors in both stages of the analysis to give a correct confidence interval. Additionally, we determined the type of outcome investigated (continuous, binary, time-to-event) and for binary outcomes what kind of target parameter was estimated (risk difference, odds ratio, relative risk, probit coefficient). We also determined whether a statistical test was used to compare the IV estimate to the 'conventional' estimate of the effect of the exposure on the outcome, what type of genetic instrument was used (single SNP or allele, multiple SNPs in separate analyses, multiple SNPs in a single analysis, combinations of SNPs e.g. haplotypes or a genetic risk score) and for those studies which used multiple SNPs in a single analysis, whether weak instrument bias was discussed. In the IV studies within one study population we also determined whether the genetic variant used as an instrument was identified or selected in the same population or if the weights for a weighted genetic risk score were derived in the same population.
2. For studies comparing the observed and expected genotype-outcome association we assessed whether the method used to obtain a point estimate of the expected genotype-outcome association was described. If the description was such that calculation of this point estimate should be possible using the data provided, we assessed whether the point estimate corresponded to our calculations (only in those studies within one population). Further, we assessed whether a confidence interval for the expected genotype-outcome association was reported, whether the method used to obtain this confidence interval was described, and whether the confidence interval incorporated the variance of both the genotype-exposure association and the exposure-outcome association.

Results

Our search returned 1911 hits, of which 594 hits remained after exclusion of conference abstracts and duplications. After reviewing the title and abstract and if necessary the fulltext article, a further 415 records were excluded for reasons listed in the flowchart in Figure 2, resulting in 179 eligible Mendelian randomisation studies. Of these 179 studies, 99 studies used data from a single study population for their main analyses,[10-59,60-108] and 80 studies used data from more than one study population (Table 1).[109-158,159-188]The included studies were published between May 2005 and

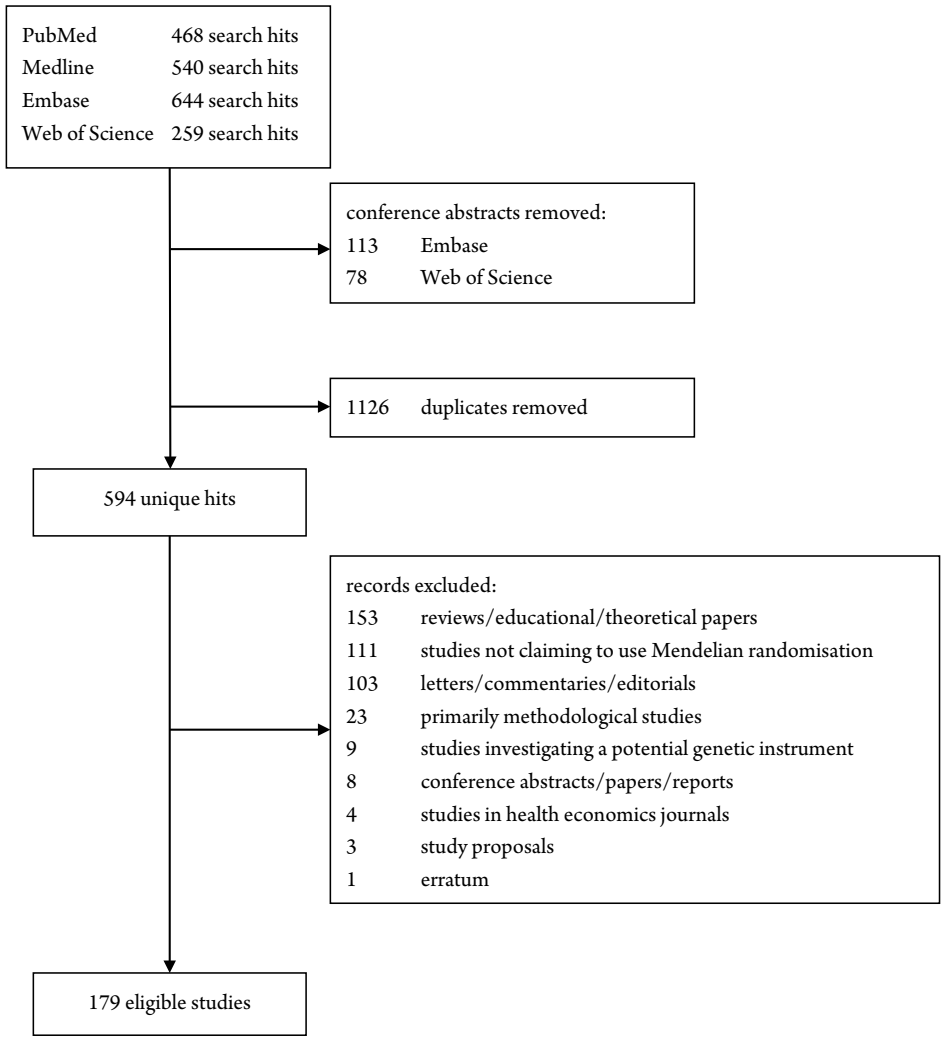


Figure 2. *Summary of literature search.*

Table 1 *Approaches used in Mendelian randomisation studies*

Data from 1 study population	(n=99)	Refs
1. Genotype used as a proxy for exposure, without further estimation [a]	38	10-47
2. Comparison of observed and expected genotype-outcome association	10	48-57
3. IV analysis[b]	48	58-105
4. Comparison of observed and expected genotype-outcome association and IV analysis	1	106
1. Unclear	2	107; 108
Data from more than 1 study population	**(n=80)**	
1. Data pooled, then analysed		
a. Genotype used as a proxy for exposure, without further estimation	2	109; 110
b. Comparison of observed and expected genotype-outcome association [a]	3	111-113
c. IV analysis[b]	14	114-127
d. Comparison of observed and expected genotype-outcome association and IV analysis	7	128-134
2. IV analyses, then meta-analysis	10	135-144
3. Meta-analysis using genotype as a proxy for exposure, without further estimation	15	134-136; 145-157
4. Meta-analyses*, followed by comparison of observed and expected genotype-outcome association[c]	13	139; 144; 157-167
5. Meta-analyses*, followed by a Wald-type/ratio estimate	9	168-176
6. Data analysed and reported separately for more than 1 population		
a. Genotype used as a proxy for exposure, without further estimation[c]	3	177; 179; 180
b. IV analysis	3	178; 181; 182
7. Multivariate meta-analysis	2	155; 183
8. Bayesian meta-analysis	1	184
9. Separate study IV-analysis	1	185
10. Meta-analysis of gene-exposure association, then ratio estimate, then meta-analysis	1	186
11. Other/unclear**	2	187; 188

Some studies used multiple approaches in non-identical sets of study populations.

[a] Two studies also performed an IV analysis for which it was unclear how the data were combined.[112;113]

[b] One study performed some of the analyses in a single study population.[118]

[c] Two studies also performed an IV analysis in a single study population.[158;177]

*Meta-analyses of genotype-exposure, exposure-outcome and/or genotype-outcome associations.

** One study first investigated the genotype-outcome association and then performed further analyses for which the approach was unclear.[187] One study used a "likelihood-based method for combining summarised genetic association estimates".[188]

December 2013. An overview of the exposures studied and the genetic instruments used is presented in Supplementary Table 1. The most frequently studied exposures were C-reactive protein (29 studies) and adiposity measures such as body mass index, fat mass and percentage body fat (25 studies).

Of the 99 studies which used data from a single study population, 38 studies (38%) used the genetic information as a proxy for the exposure by investigating the genotype-outcome association without further estimation of either the causal effect of the exposure on the outcome or of the expected genotype-outcome association (Table 1). Forty-eight studies (48%) used IV analysis to estimate the effect of genetically determined variation in exposure levels on the outcome. Ten studies compared the observed association between the genotype and the outcome with the expected association based on the genotype-exposure association and the exposure-outcome association. One study used both these latter two approaches. For two studies we could not categorise the methods used into any of the aforementioned approaches.

Of the 80 studies which used data from multiple study populations, 26 (33%) studies pooled the data from the different studies and subsequently analysed the pooled data (Table 1). Ten studies performed an IV analysis in the different studies followed by a meta-analysis. Forty-one studies (51%) first performed a meta-analysis of one or more of the genotype-exposure, exposure-outcome and genotype-outcome associations, 26 of which subsequently used these meta-analysed associations for further estimation of either the causal effect of the exposure on the outcome or of the expected genotype-outcome association. In total, 52 studies (65%) used some form of IV analysis to obtain a causal effect of the exposure on the outcome. A further 23 studies compared the observed and expected genotype-outcome associations.

Table 2 summarises the reporting of the Mendelian randomisation assumptions. Reporting of assumptions was assessed in 178 studies, because the design of one study was so different from the general Mendelian randomisation design that the assumptions could not be assessed. A total of 37 out of 98 studies (38%) which used a single study population and 42 out of 80 studies (53%) which used multiple study populations explicitly discussed the plausibility of all three Mendelian randomisation assumptions in the context of their study.

Among the studies which performed an IV analysis, those using a single study population most frequently studied a continuous outcome, whereas those using multiple study populations most frequently studied a binary outcome and estimated an odds ratio (Table 3). The statistical methods used in these formal IV studies are

shown in Table 4. Two-stage least squares (2-SLS) regression was the most common method used in studies within one study population (n=26, 53%). Ten studies within multiple study populations also used this method. One study used 2-SLS with a binary outcome, but it did not mention whether heteroskedasticity robust standard errors were used.[69] Among the studies which used multiple study populations a Wald-type or ratio estimator was most frequently used (n=16, 31%). The method used to obtain the confidence interval for the ratio estimate was a Taylor series expansion (termed the delta method[138,141,168,169,186] or Taylor expansion[170]), Fieller method,[120,176-178] or was not described. Three studies in a single study population also used a Wald-type/ratio estimator, but two of these studies did not report a confidence interval. Other

Table 2. *Reporting of Mendelian randomisation assumptions.*

Criterium	1 Study population (n=98)	>1 Study population (n=80)
Strength of genetic instrument-exposure association (assumption 1)		
Verified in data using F-statistic	33	26
Otherwise verified in data (e.g. using risk difference or odds ratio)	53	45
Reported from literature	4	4
Not reported	8	5‡
Plausibility of exclusion restriction assumption discussed (assumption 2)	56	55
Independence assumption (assumption 3)		
Instrument-confounder associations shown & assumption further discussed theoretically†	20	16
Instrument-confounder associations shown, assumption not further discussed	30	21
Investigation of instrument-confounder associations mentioned, not shown & assumption further discussed theoretically	4	0
Investigation of instrument-confounder associations mentioned, not shown & assumption not further discussed	8	0
Plausibility of assumption theoretically discussed only	7	18
Plausibility of assumption not discussed	29	25

*Reporting of assumptions was not assessed in one study, because its design was vastly different from the general design of a Mendelian randomisation study. The total number of studies within 1 study population is therefore 98.

†Potential association with unmeasured confounders discussed and/or population stratification discussed.

‡Two studies reported a p-value only.

methods used were control functions (n=8 in total), IV probit regression (n=4), generalised method of moments (n=8), generalised least squares regression (n=5), quasi-likelihood and variance function (n=4) and a two stage approach with a linear first stage and a logistic second stage (n=5). Four of the studies which used this last approach did not report how the correct confidence interval was obtained,[125,133,135,158] and the fifth used a sandwich estimator.[114] The IV method was insufficiently described in fourteen studies. In six of these studies there was a discrepancy between the statistical method reportedly used (2-SLS) and the effect estimate reported (OR).[100,128-131,134] Another study seemingly did not take into account the variance of the genotype-exposure association in the variance of the IV estimate, which would result in too narrow a confidence interval.[101]

Of the 101 studies which used one of the approaches which yields an IV estimate, 48 reported tests of the difference between the IV estimate and the conventional estimate: the most commonly used were (a variant of) the Durbin-Wu-Hausman test (29 studies),[58-60,62,64-66,68,70-72,74-77,79-87,94,98,105,116,139] and (a variant of) the Bland-Altman test (10 studies).[112,113,117,119,128-132,134] The types of genetic instrument used (e.g. a single SNP or a genetic risk score) in the IV analysis studies are listed in Supplementary table 2. Of the 25 studies which used multiple SNPs in a single analysis 13 mentioned

Table 3. *Types of outcome and parameters estimated in IV Mendelian randomisation studies.*

Type of outcome	1 Study population (n=49)	Refs	>1 Study population* (n=52)	Refs
Continuous	37	58-68; 70-89; 94; 98; 103-106	14	114-116; 118; 126;135-139; 168; 170; 182; 185
Binary				
Risk difference	3	69; 93; 99	0	-
Relative risk	2	81; 102	2	118; 124
Odds ratio	7	67; 82; 88; 90-92; 100	37	112-114; 117; 119-121; 123; 125; 128-135; 139-144; 155; 158; 168; 169; 171-176; 178; 181; 183; 184; 186
Probit coefficient	1	83	1	122
Time-to-event	4	95-97; 101	5	124; 138; 168; 177; 178
Unclear	0	-	1	127

The total number of types of outcome and parameters estimated exceeds the total number of studies because some studies included multiple types of outcomes.

weak instrument bias, with two studies very specifically discussing it in relation to using multiple instruments.[117,185] Of the 49 studies which used IV methods and were performed in one study population, 14 evidently identified or selected the genetic variant used as an instrument in the same population or derived weights for a weighted genetic risk score in the same population.[61,62,65,66,70,71,82,85,87,91,93,102,104,105]

Table 4 *Statistical methods used in the instrumental variable studies.*

Method	1 Study population (n=49)	Refs	>1 Study population* (n=52)	Refs
Two-stage least squares	26	58-83	10	114-116; 118; 135-139; 182
Instrumental variable regression in Stata, not further specified (2-SLS, GMM or LIML)	5	84-88	0	-
Control function	6	81; 82; 89-92	2	139; 143
Instrumental variable probit regression	3	67; 83; 93	1	122
GMM	2	94;98	0	-
Multiplicative GMM	0	-	6	117-121; 124
Generalised least squares regression	1	95	4	112; 113; 123; 132
Two-stage: linear first stage, logistic second stage	0	-	5	114; 125; 133; 135; 158
Quasi-likelihood and variance function	1	88	3	140; 144; 181
Ratio/Wald-type estimator	1	99	17	120; 124; 138; 141; 142; 168-178; 186
Ratio/Wald-type estimator without confidence interval	2	96;97	0	-
Insufficiently described/unclear	7	100-106	7	126-131; 134
Other**	0	-	4	155; 183-185

The total number of statistical methods exceeds the total number of studies because some studies investigated multiple statistical methods.
Abbreviations: 2-SLS, two-stage least squares; GMM, generalised method of moments; LIML, limited-information maximum likelihood.
*Including the two studies which used multiple study populations, but performed the IV analysis in a single study population.
** See Table 1.

In 3 of the 11 studies comparing the observed gene-outcome association to the expected gene-outcome association in one study population we could not reconstruct the point estimate of the expected association from the data.[48,50,57] Four studies did not report a confidence interval for the expected genotype-outcome association.[52,54,56,106] In a further five studies the methods used to calculate this confidence interval were unclear,[48-50,53,57] and in one study only the error in the exposure-outcome association seemed to have been taken into account in the calculation of this confidence interval.[55] Only one study adequately described the methods used to obtain the point estimate and confidence interval (bootstrapping) of the expected genotype-outcome association.[51] In the 23 studies which employed this approach using more than one study population, three only took into account the error in the exposure-outcome association and not the error in the genotype-exposure association,[159,164,165] and 16 studies did not describe how the confidence interval was obtained.

Discussion

Most Mendelian randomisation studies either performed some form of IV analysis (49% of studies within 1 study population and 65% of studies within multiple study populations) or used the genotype as a proxy for the exposure without further estimation. A third approach used less frequently was to compare the observed genotype-outcome association to the expected genotype-outcome association. Although validity of the three main Mendelian randomisation assumptions is required regardless of the approach used, only 44% of studies adequately discussed the plausibility of these assumptions. The methods used to obtain an IV estimate were not always adequately described. For those studies which are performed using multiple study populations, the range of approaches used was very broad, because of further differentiation according to the way the data from the different studies were combined. Here we will discuss our findings and propose recommendations for the reporting of Mendelian randomisation studies.

To our knowledge there is one paper which previously reviewed MR studies, which included a much smaller number of studies. Its main focus was on whether the Mendelian randomisation studies reported results that were compatible with a causal association, which was the case for over half of their reviewed studies.[189] In contrast, our review focussed on the approach used and on the discussion of the assumptions and the reporting of the statistical methods used. The previous review also noted that many studies applied IV analysis to a binary outcome, using methods which had not quite been validated,[189] which is an issue which we will also discuss later.

Our meta-epidemiological study has several limitations. With respect to study selection, we investigated what methods were used in studies *stating* that they used Mendelian randomisation or that they used a genetic IV. Importantly, we were unable to include studies which apply the same principles without using the term Mendelian randomisation or genetic IV because these could not feasibly be found using a systematic search strategy. We do not know to what extent our results apply to these studies, but suspect the discussion of Mendelian randomisation assumptions in particular is likely to be insufficient in many of these studies. Importantly, the focus of our review was on the quality of *reporting* of methods used in Mendelian randomisation studies. We did not assess whether the statistical method used to obtain an IV estimate was actually appropriate. We investigated whether the statistical method used was adequately described, whether it was consistent with the estimates reported, and if any evident mistakes were made. Similarly, we focussed on whether plausibility of MR assumptions was discussed, not on whether we considered them likely to hold.

With regard to the Mendelian randomisation approach used we found that a majority of studies performed some form of IV analysis, but a substantial proportion of studies used the genotype as a proxy for the exposure without performing a formal IV analysis. This raises the question whether either of these approaches, or the third option of comparing the observed and expected genotype-outcome association should be preferred. This depends on the aim of the study: for a test of causality testing the presence of a genotype-outcome association is sufficient.[1,190] Often the aim will be a quantification of the causal effect of the exposure on the outcome. We note that IV analysis is more suited to this aim than a comparison of the observed and the expected effect of the genotype on the outcome, although some may find the latter approach more intuitive. Showing the association between the genotype (or genetic score) and the outcome is always advisable as it increases the transparency of the study by showing the data as they are. Further analyses can subsequently be undertaken.[191] When considering whether a formal IV analysis is appropriate, further aspects of the underlying biology of the genotype-phenotype association need to be taken into account to avoid misleading inferences.[192] A recent paper discusses a number of situations in which a formal IV analysis may give biased results, but a Mendelian randomisation approach looking only at the genotype-outcome association can validly be used as a test of causality.[190] Another recent paper specifically discusses smoking as an example of an exposure for which the measurement does not fully capture the underlying exposure, which gives a biased estimate of the effect of the measured exposure on the outcome if an IV analysis is performed in a Mendelian randomisation study.[4]

With regard to the discussion of the Mendelian randomisation assumptions we found that fewer than half of studies adequately discussed all three assumptions. Some studies did mention what the assumptions are and how they can be violated in general terms, but did not discuss how plausible the assumptions were for the specific setting of their study. An aspect of the assumptions which can be evaluated using the data is whether there is an association between the genetic instrument and measured confounders. This may be more difficult for studies which use multiple study populations, but an effort to obtain this information from those studies in which it is available is warranted. Among the studies which performed an IV analysis in a single study population, we identified 14 studies in which SNPs were detected or selected, or genetic risk score weights were derived in that same study population. This can bias Mendelian randomisation estimates.[116,193] The number of studies in which we found this to have occurred may be an underestimation, because some study populations are used for multiple Mendelian randomisation studies and the later studies may not report the detection of SNPs in a previous study in the same population.

With regard to the IV methods used, we found that two-stage least squares regression and a Wald-type/ratio estimator were the most commonly used methods. We also found that a considerable number of the Mendelian randomisation studies which used IV methods estimated an odds ratio or risk ratio, especially in those studies which used data from multiple study populations. However, which methods are appropriate for IV estimation of causal odds ratios or risk ratios is a methodological challenge of IV analysis that has not yet been fully resolved. Several methodological studies have investigated this issue in recent years.[194-198] One of the reviewed MR studies mentioned that the Wald-type estimator used to estimate an odds ratio was an approximate method.[169] The properties and limitations of these IV methods used to estimate a causal odds ratio deserve more attention in the Mendelian randomisation studies in which they are used.

Overall, we conclude from our review the standard of reporting of Mendelian randomisation studies should be improved. Existing guidelines and recommendations for the reporting of IV analyses largely apply to Mendelian randomisation studies (the extent depending on the Mendelian randomisation approach used).[7,8] In addition to these recommendations we have formulated a checklist of Mendelian randomisation-specific reporting recommendations in Box 1.

In conclusion, studies stating that they perform a Mendelian randomisation study within one study population broadly fall into three categories: studies using a genotype as a proxy for exposure without further estimation, studies performing IV analysis

using a genotype as an instrument and studies comparing observed and expected genotype-outcome associations. Plausibility of underlying Mendelian randomisation assumptions are not always discussed, but as these assumptions are crucial for validity of MR studies, they should always be discussed in the specific context of the study. If IV methods are used to estimate a causal effect of the exposure, the statistical methods used should be clearly explained. Studies using data from multiple populations should also clearly report how data or estimates are combined.

Reference List

(1) Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med 2008 Apr 15;27(8):1133-63.

(2) Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol 2003 Feb;32(1):1-22.

(3) Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. Lancet 1986 Mar 1;1(8479):507-8.

(4) Taylor AE, Davies NM, Ware JJ, VanderWeele T, Davey Smith G, Munafo MR. Mendelian randomization in health research: using appropriate genetic variants and avoiding biased estimates. Econ Hum Biol 2014 Mar;13:99-106.

(5) Sheehan NA, Meng S, Didelez V. Mendelian randomisation: a tool for assessing causality in observational epidemiology. Methods Mol Biol 2011;713:153-66.

(6) Didelez V, Meng S, Sheehan NA. Assumptions of IV Methods for Observational Epidemiology. Statist Sci 2010;22-40.

(7) Davies NM, Davey Smith G, Windmeijer F, Martin RM. Issues in the reporting and conduct of instrumental variable studies: a systematic review. Epidemiology 2013 May;24(3):363-9.

(8) Swanson SA, Hernán MA. Commentary: how to report instrumental variable analyses (suggestions welcome). Epidemiology 2013 May;24(3):370-4.

(9) Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. Epidemiology 2006 May;17(3):260-7.

(10) Sharma NK, Gupta A, Prabhakar S, et al. Association between CFH Y402H polymorphism and age related macular degeneration in North Indian cohort. PLoS One 2013;8(7):e70193.

(11) Shaheen SO, Rutterford C, Zuccolo L, et al. Prenatal alcohol exposure and childhood atopic disease: a Mendelian randomization approach. J Allergy Clin Immunol 2014 Jan;133(1):225-32.

(12) Humphriss R, Hall A, May M, Zuccolo L, Macleod J. Prenatal alcohol exposure and childhood balance ability: findings from a UK birth cohort study. BMJ Open 2013;3(6).

(13) Bonilla C, Lawlor DA, Taylor AE, et al. Vitamin B-12 status during pregnancy and child's IQ at age 8: a Mendelian randomization study in the Avon longitudinal study of parents and children. PLoS One 2012;7(12):e51084.

(14) Bonilla C, Lawlor DA, Ben-Shlomo Y, et al. Maternal and offspring fasting glucose and type 2 diabetes-associated genetic variants and cognitive function at age 8: a Mendelian randomization study in the Avon Longitudinal Study of Parents and Children. BMC Med Genet 2012;13:90.

(15) Bonilla C, Gilbert R, Kemp JP, et al. Using genetic proxies for lifecourse sun exposure to assess the causal relationship of sun exposure with circulating vitamin d and prostate cancer risk. Cancer Epidemiol Biomarkers Prev 2013 Apr;22(4):597-606.

(16) Alegret JM, Aragonès G, Elosua R, et al. The relevance of the association between inflammation and atrial fibrillation. Eur J Clin Invest 2013 Apr;43(4):324-31.

(17)) Almon R, Álvarez-León EE, Serra-Majem L. Association of the European lactase persistence variant (LCT-13910 C>T polymorphism) with obesity in the Canary Islands. PLoS One 2012;7(8):e43978.
(18) Bjørngaard JH, Gunnell D, Elvestad MB, et al. The causal role of smoking in anxiety and depression: a Mendelian randomization analysis of the HUNT study. Psychol Med 2013 Apr;43(4):711-9.
(19) Attermann J, Obel C, Bilenberg N, Nordenbæk CM, Skytthe A, Olsen J. Traits of ADHD and autism in girls with a twin brother: a Mendelian randomization study. Eur Child Adolesc Psychiatry 2012 Sep;21(9):503-9.
(20) Yang Q, Bailey L, Clarke R, et al. Prospective study of methylenetetrahydrofolate reductase (MTHFR) variant C677T and risk of all-cause and cardiovascular disease mortality among 6000 US adults. Am J Clin Nutr 2012 May;95(5):1245-53.
(21) van Durme YM, Lahousse L, Verhamme KM, et al. Mendelian randomization study of interleukin-6 in chronic obstructive pulmonary disease. Respiration 2011;82(6):530-8.
(22) Scott JA, Berkley JA, Mwangi I, et al. Relation between falciparum malaria and bacteraemia in Kenyan children: a population-based, case-control study and a longitudinal study. Lancet 2011 Oct 8;378(9799):1316-23.
(23) Chmielewski M, Verduijn M, Drechsler C, et al. Low cholesterol in dialysis patients--causal factor for mortality or an effect of confounding? Nephrol Dial Transplant 2011 Oct;26(10):3325-31.
(24) Bolton CE, Schumacher W, Cockcroft JR, et al. The CRP genotype, serum levels and lung function in men: the Caerphilly Prospective Study. Clin Sci (Lond) 2011 Apr;120(8):347-55.
(25) Kröger J, Zietemann V, Enzenbach C, et al. Erythrocyte membrane phospholipid fatty acids, desaturase activity, and dietary fatty acids in relation to risk of type 2 diabetes in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam Study. Am J Clin Nutr 2011 Jan;93(1):127-42.
(26) Welsh P, Polisecki E, Robertson M, et al. Unraveling the directional link between adiposity and inflammation: a bidirectional Mendelian randomization approach. J Clin Endocrinol Metab 2010 Jan;95(1):93-9.
(27) Trompet S, Jukema JW, Katan MB, et al. Apolipoprotein e genotype, plasma cholesterol, and cancer: a Mendelian randomization study. Am J Epidemiol 2009 Dec 1;170(11):1415-21.
(28) Almon R, Álvarez-León EE, Engfeldt P, Serra-Majem L, Magnuson A, Nilsson TK. Associations between lactase persistence and the metabolic syndrome in a cross-sectional study in the Canary Islands. Eur J Nutr 2010 Apr;49(3):141-6.
(29) Brennan P, McKay J, Moore L, et al. Obesity and cancer: Mendelian randomization approach utilizing the FTO genotype. Int J Epidemiol 2009 Aug;38(4):971-5.
(30) Drenos F, Talmud PJ, Casas JP, et al. Integrated associations of genotypes with multiple blood biomarkers linked to coronary heart disease risk. Hum Mol Genet 2009 Jun 15;18(12):2305-16.
(31) Lim LS, Tai ES, Aung T, et al. Relation of age-related cataract with obesity and obesity genes in an Asian population. Am J Epidemiol 2009 May 15;169(10):1267-74.
(32) Giltay EJ, van Reedt Dortland AK, Nissinen A, et al. Serum cholesterol, apolipoprotein E genotype and depressive symptoms in elderly European men: the FINE study. J Affect Disord 2009 Jun;115(3):471-7.

(33) Irons DE, McGue M, Iacono WG, Oetting WS. Mendelian randomization: a novel test of the gateway hypothesis and models of gene-environment interplay. Dev Psychopathol 2007;19(4):1181-95.
(34) Herder C, Klopp N, Baumert J, et al. Effect of macrophage migration inhibitory factor (MIF) gene variants and MIF serum concentrations on the risk of type 2 diabetes: results from the MONICA/KORA Augsburg Case-Cohort Study, 1984-2002. Diabetologia 2008 Feb;51(2):276-84.
(35) Keavney B, Danesh J, Parish S, et al. Fibrinogen and coronary heart disease: test of causality by 'Mendelian randomization'. Int J Epidemiol 2006 Aug;35(4):935-43.
(36) Bech BH, Autrup H, Nohr EA, Henriksen TB, Olsen J. Stillbirth and slow metabolizers of caffeine: comparison by genotypes. Int J Epidemiol 2006 Aug;35(4):948-53.
(37) Zuccolo L, Lewis SJ, Davey Smith G, et al. Prenatal alcohol exposure and offspring cognition and school performance. A 'Mendelian randomization' natural experiment. International Journal of Epidemiology 2013 Oct;42(5):1358-70.
(38) Aramini B, Kim C, Diangelo S, et al. Donor surfactant protein D (SP-D) polymorphisms are associated with lung transplant outcome. American Journal of Transplantation 2013 Aug;13(8):2130-6.
(39) Yarwood A, Martin P, Bowes J, et al. Enrichment of vitamin D response elements in RA-associated loci supports a role for vitamin D in the pathogenesis of RA. Genes and Immunity 2013 Jul;14(5):325-9.
(40) Travis RC, Appleby PN, Siddiq A, et al. Genetic variation in the lactase gene, dairy product intake and risk for prostate cancer in the European prospective investigation into cancer and nutrition. International Journal of Cancer 2013 Apr 15;132(8):1901-10.
(41) Veen G, Giltay EJ, Van Vliet IM, et al. C-reactive protein polymorphisms are associated with the cortisol awakening response in basal conditions in human subjects. Stress 2011 Mar;14(2):128-35.
(42) Koshy B, Miyashita A, St.Jean P, et al. Genetic deficiency of plasma lipoprotein-associated phospholipase A 2 (PLA2G7 V297F null mutation) and risk of Alzheimer's disease in Japan. Journal of Alzheimer's Disease 2010;21(3):775-80.
(43) Pierce BL, Ahsan H. Genetic susceptibility to type 2 diabetes is associated with reduced prostate cancer risk. Human Heredity 2010 Mar;69(3):193-201.
(44) Thuesen BH, Husemoen LLN, Fenger M, Linneberg A. Lack of association between the MTHFR (C677T) polymorphism and atopic disease. Clinical Respiratory Journal 2009 Apr;3(2):102-8.
(45) Granell R, Heron J, Lewis S, Davey Smith G, Sterne JAC, Henderson J. The association between mother and child MTHFR C677T polymorphisms, dietary folate intake and childhood atopy in a population-based, longitudinal birth cohort. Clinical and Experimental Allergy 2008 Feb;38(2):320-8.
(46) Heidrich J, Wellmann J, Doring A, Illig T, Keil U. Alcohol consumption, alcohol dehydrogenase and risk of coronary heart disease in the MONICA/KORA-Augsburg cohort 1994/1995-2002. European Journal of Cardiovascular Prevention and Rehabilitation 2007 Dec;14(6):769-74.
(47) Almeida OP, Hankey GJ, Yeap BB, Golledge J, Flicker L. The triangular association of ADH1B genetic polymorphism, alcohol consumption and the risk of depression in older men. Mol Psychiatry 2013 Sep 10.

(48) Dai X, Yuan J, Yao P, et al. Association between serum uric acid and the metabolic syndrome among a middle- and old-age Chinese population. Eur J Epidemiol 2013 Aug;28(8):669-76.
(49) Yao WM, Zhang HF, Zhu ZY, et al. Genetically elevated levels of circulating triglycerides and brachial-ankle pulse wave velocity in a Chinese population. J Hum Hypertens 2013 Apr;27(4):265-70.
(50) Gan W, Guan Y, Wu Q, et al. Association of TMPRSS6 polymorphisms with ferritin, hemoglobin, and type 2 diabetes risk in a Chinese Han population. Am J Clin Nutr 2012 Mar;95(3):626-32.
(51) Breitling LP, Koenig W, Fischer M, et al. Type II secretory phospholipase A2 and prognosis in patients with stable coronary heart disease: mendelian randomization study. PLoS One 2011;6(7):e22318.
(52) Rasmussen-Torvik LJ, Li M, Kao WH, et al. Association of a fasting glucose genetic risk score with subclinical atherosclerosis: The Atherosclerosis Risk in Communities (ARIC) study. Diabetes 2011 Jan;60(1):331-5.
(53) Wu Y, Li H, Loos RJ, et al. RBP4 variants are significantly associated with plasma RBP4 levels and hypertriglyceridemia risk in Chinese Hans. J Lipid Res 2009 Jul;50(7):1479-86.
(54) Di Paola R, Marucci A, Fontana A, et al. Role of obesity on all-cause mortality in whites with type 2 diabetes from Italy. Acta Diabetologica 2013 Dec;50(6):971-6.
(55) Menzaghi C, De Cosmo S, Copetti M, et al. Relationship between ADIPOQ gene, circulating high molecular weight adiponectin and albuminuria in individuals with normal kidney function: Evidence from a family-based study. Diabetologia 2011 Apr;54(4):812-8.
(56) Oei L, Campos-Obando N, Dehghan A, et al. Dissecting the relationship between high-sensitivity serum C-reactive protein and increased fracture risk: the Rotterdam Study. Osteoporos Int 2014 Apr;25(4):1247-54.
(57) Tian Q, Jia J, Ling S, Liu Y, Yang S, Shao Z. A causal role for circulating miR-34b in osteosarcoma. Eur J Surg Oncol 2014 Jan;40(1):67-72.
(58) Bouthoorn SH, van Lenthe FJ, Kiefte-de Jong JC, et al. Genetic taste blindness to bitter and body composition in childhood: a Mendelian randomization design. Int J Obes (Lond) 2014 Jul;38(7):1005-10.
(59) Lee HA, Park EA, Cho SJ, et al. Mendelian randomization analysis of the effect of maternal homocysteine during pregnancy, as represented by maternal MTHFR C677T genotype, on birth weight. J Epidemiol 2013 Sep 5;23(5):371-5.
(60) Warodomwichit D, Sritara C, Thakkinstian A, et al. Causal inference of the effect of adiposity on bone mineral density in adults. Clin Endocrinol (Oxf) 2013 May;78(5):694-9.
(61) Jensen MK, Bartz TM, Djousse L, et al. Genetically elevated fetuin-A levels, fasting glucose levels, and risk of type 2 diabetes: the cardiovascular health study. Diabetes Care 2013 Oct;36(10):3121-7.
(62) Gao H, Fall T, van Dam RM, et al. Evidence of a causal relationship between adiponectin levels and insulin sensitivity: a Mendelian randomization study. Diabetes 2013 Apr;62(4):1338-44.

(63) Alwan NA, Lawlor DA, McArdle HJ, Greenwood DC, Cade JE. Exploring the relationship between maternal iron status and offspring's blood pressure and adiposity: a Mendelian randomization study. Clin Epidemiol 2012;4:193-200.
(64) Oikonen M, Wendelin-Saarenhovi M, Lyytikäinen LP, et al. Associations between serum uric acid and markers of subclinical atherosclerosis in young adults. The cardiovascular risk in Young Finns study. Atherosclerosis 2012 Aug;223(2):497-503.
(65) Lyngdoh T, Vuistiner P, Marques-Vidal P, et al. Serum uric acid and adiposity: deciphering causality using a bidirectional Mendelian randomization approach. PLoS One 2012;7(6):e39321.
(66) Guessous I, Dobrinas M, Kutalik Z, et al. Caffeine intake and CYP1A2 variants associated with high caffeine intake protect non-smokers from hypertension. Hum Mol Genet 2012 Jul 15;21(14):3283-92.
(67) Au Yeung SL, Jiang C, Cheng KK, et al. Moderate alcohol use and cardiovascular disease from Mendelian randomization. PLoS One 2013;8(7):e68054.
(68) Au Yeung SL, Jiang CQ, Cheng KK, et al. Evaluation of moderate alcohol use and cognitive function among men using a Mendelian randomization design in the Guangzhou biobank cohort study. Am J Epidemiol 2012 May 15;175(10):1021-8.
(69) Lewis SJ, Araya R, Davey Smith G, et al. Smoking is associated with, but does not cause, depressed mood in pregnancy--a mendelian randomization study. PLoS One 2011;6(7):e21689.
(70) Conen D, Vollenweider P, Rousson V, et al. Use of a Mendelian randomization approach to assess the causal relation of gamma-Glutamyltransferase with blood pressure and serum insulin levels. Am J Epidemiol 2010 Dec 15;172(12):1431-41.
(71) Bochud M, Marquant F, Marques-Vidal PM, et al. Association between C-reactive protein and adiposity in women. J Clin Endocrinol Metab 2009 Oct;94(10):3969-77.
(72) Timpson NJ, Sayers A, Davey Smith G, Tobias JH. How does body fat influence bone mass in childhood? A Mendelian randomization approach. J Bone Miner Res 2009 Mar;24(3):522-33.
(73) Kivimäki M, Lawlor DA, Davey Smith G, et al. Does high C-reactive protein concentration increase atherosclerosis? The Whitehall II Study. PLoS One 2008;3(8):e3013.
(74) Brunner EJ, Kivimäki M, Witte DR, et al. Inflammation, insulin resistance, and diabetes--Mendelian randomization using CRP haplotypes points upstream. PLoS Med 2008 Aug 12;5(8):e155.
(75) Lawlor DA, Timpson NJ, Harbord RM, et al. Exploring the developmental overnutrition hypothesis using parental-offspring associations and FTO as an instrumental variable. PLoS Med 2008 Mar 11;5(3):e33.
(76) Viikari LA, Huupponen RK, Viikari JS, et al. Relationship between leptin and C-reactive protein in young Finnish adults. J Clin Endocrinol Metab 2007 Dec;92(12):4753-8.
(77) Timpson NJ, Lawlor DA, Harbord RM, et al. C-reactive protein and its role in metabolic syndrome: mendelian randomisation study. Lancet 2005 Dec 3;366(9501):1954-9.
(78) Binder AM, Michels KB. The causal effect of red blood cell folate on genome-wide methylation in cord blood: a Mendelian randomization approach. BMC Bioinformatics 2013;14:353.
(79) Debette S, Wolf C, Lambert JC, et al. Abdominal obesity and lower gray matter volume: a Mendelian randomization study. Neurobiol Aging 2014 Feb;35(2):378-86.

(80) Thakkinstian A, Chailurkit L, Warodomwichit D, et al. Causal relationship between body mass index and fetuin-A level in the asian population: a bidirectional mendelian randomization study. Clin Endocrinol (Oxf) 2014 Aug;81(2):197-203.
(81) Haring R, Teumer A, Völker U, et al. Mendelian randomization suggests non-causal associations of testosterone with cardiometabolic risk factors and mortality. Andrology 2013 Jan;1(1):17-23.
(82) Islam M, Jafar TH, Wood AR, et al. Multiple genetic variants explain measurable variance in type 2 diabetes-related traits in Pakistanis. Diabetologia 2012 Aug;55(8):2193-204.
(83) Kivimäki M, Jokela M, Hamer M, et al. Examining overweight and obesity as risk factors for common mental disorders using fat mass and obesity-associated (FTO) genotype-instrumented analysis: The Whitehall II Study, 1985-2004. Am J Epidemiol 2011 Feb 15;173(4):421-9.
(84) Jokela M, Elovainio M, Keltikangas-Järvinen L, et al. Body mass index and depressive symptoms: instrumental-variables regression with genetic risk score. Genes Brain Behav 2012 Sep 7.
(85) Kivimäki M, Magnussen CG, Juonala M, et al. Conventional and Mendelian randomization analyses suggest no association between lipoprotein(a) and early atherosclerosis: the Young Finns Study. Int J Epidemiol 2011 Apr;40(2):470-8.
(86) Kivimäki M, Davey Smith G, Timpson NJ, et al. Lifetime body mass index and later atherosclerosis risk in young adults: examining causal links using Mendelian randomization in the Cardiovascular Risk in Young Finns study. Eur Heart J 2008 Oct;29(20):2552-60.
(87) Frayling TM, Rafiq S, Murray A, et al. An interleukin-18 polymorphism is associated with reduced serum concentrations and better physical functioning in older people. J Gerontol A Biol Sci Med Sci 2007 Jan;62(1):73-8.
(88) Davey Smith G, Lawlor DA, Harbord R, et al. Association of C-reactive protein with blood pressure and hypertension: life course confounding and mendelian randomization tests of causality. Arterioscler Thromb Vasc Biol 2005 May;25(5):1051-6.
(89) Lawlor DA, Nordestgaard BG, Benn M, Zuccolo L, Tybjærg-Hansen A, Davey Smith G. Exploring causal associations between alcohol and coronary heart disease risk factors: findings from a Mendelian randomization study in the Copenhagen General Population Study. Eur Heart J 2013 Aug;34(32):2519-28.
(90) Theodoratou E, Palmer T, Zgaga L, et al. Instrumental variable estimation of the causal effect of plasma 25-hydroxy-vitamin D on colorectal cancer risk: a mendelian randomization analysis. PLoS One 2012;7(6):e37662.
(91) Collin SM, Metcalfe C, Palmer TM, et al. The causal roles of vitamin B(12) and transcobalamin in prostate cancer: can Mendelian randomization analysis provide definitive answers? Int J Mol Epidemiol Genet 2011;2(4):316-27.
(92) Lawlor DA, Harbord RM, Tybjærg-Hansen A, et al. Using genetic loci to understand the relationship between adiposity and psychological distress: a Mendelian Randomization study in the Copenhagen General Population Study of 53,221 adults. J Intern Med 2011 May;269(5):525-37.
(93) Mumby HS, Elks CE, Li S, et al. Mendelian Randomisation Study of Childhood BMI and Early Menarche. J Obes 2011;2011:180729.

(94) Timpson NJ, Nordestgaard BG, Harbord RM, et al. C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. Int J Obes (Lond) 2011 Feb;35(2):300-8.
(95) Zacho J, Tybjærg-Hansen A, Nordestgaard BG. C-reactive protein and all-cause mortality--the Copenhagen City Heart Study. Eur Heart J 2010 Jul;31(13):1624-32.
(96) Verduijn M, Prein RA, Stenvinkel P, et al. Is fetuin-A a mortality risk factor in dialysis patients or a mere risk marker? A Mendelian randomization approach. Nephrol Dial Transplant 2011 Jan;26(1):239-45.
(97) Fisher E, Stefan N, Saar K, et al. Association of AHSG gene polymorphisms with fetuin-A plasma levels and cardiovascular diseases in the EPIC-Potsdam study. Circ Cardiovasc Genet 2009 Dec;2(6):607-13.
(98) Timpson NJ, Harbord R, Davey Smith G, Zacho J, Tybjærg-Hansen A, Nordestgaard BG. Does greater adiposity increase blood pressure and hypertension risk?: Mendelian randomization using the FTO/MC4R genotype. Hypertension 2009 Jul;54(1):84-90.
(99) Kang H, Kreuels B, Adjei O, Krumkamp R, May J, Small DS. The causal effect of malaria on stunting: a Mendelian randomization and matching approach. International Journal of Epidemiology 2013 Oct;42(5):1390-8.
(100) Qin XY, Tian J, Fang K, et al. [Mendelian randomization study of the relationship between high-density lipoprotein cholesterol and age-related macular degeneration]. Beijing Da Xue Xue Bao 2012 Jun 18;44(3):407-11.
(101) Trummer O, Pilz S, Hoffmann MM, et al. Vitamin D and mortality: a Mendelian randomization study. Clin Chem 2013 May;59(5):793-7.
(102) You NC, Chen BH, Song Y, et al. A prospective study of leukocyte telomere length and risk of type 2 diabetes in postmenopausal women. Diabetes 2012 Nov;61(11):2998-3004.
(103) McArdle PF, Whitcomb BW, Tanner K, Mitchell BD, Shuldiner AR, Parsa A. Association between bilirubin and cardiovascular disease risk factors: using Mendelian randomization to assess causal inference. BMC Cardiovasc Disord 2012;12:16.
(104) Parsa A, Brown E, Weir MR, et al. Genotype-based changes in serum uric acid affect blood pressure. Kidney Int 2012 Mar;81(5):502-7.
(105) Sunyer J, Pistelli R, Plana E, et al. Systemic inflammation, genetic susceptibility and lung function. Eur Respir J 2008 Jul;32(1):92-7.
(106) Mente A, Meyre D, Lanktree MB, et al. Causal relationship between adiponectin and metabolic traits: a Mendelian randomization study in a multiethnic population. PLoS One 2013;8(6):e66808.
(107) Love-Gregory L, Sherva R, Schappe T, et al. Common CD36 SNPs reduce protein expression and may contribute to a protective atherogenic profile. Hum Mol Genet 2011 Jan 1;20(1):193-201.
(108) Nagele P, Zeugswetter B, Wiener C, et al. Influence of methylenetetrahydrofolate reductase gene polymorphisms on homocysteine concentrations after nitrous oxide anesthesia. Anesthesiology 2008 Jul;109(1):36-43.
(109) Klovaite J, Nordestgaard BG, Tybjærg-Hansen A, Benn M. Elevated fibrinogen levels are associated with risk of pulmonary embolism, but not with deep venous thrombosis. Am J Respir Crit Care Med 2013 Feb 1;187(3):286-93.

(110) Rius-Ottenheim N, de Craen AJ, Geleijnse JM, et al. C-reactive protein haplotypes and dispositional optimism in obese and nonobese elderly subjects. Inflamm Res 2012 Jan;61(1):43-51.
(111) Stender S, Frikke-Schmidt R, Nordestgaard BG, Tybjærg-Hansen A. Extreme bilirubin levels as a causal risk factor for symptomatic gallstone disease. JAMA Intern Med 2013 Jul 8;173(13):1222-8.
(112) Dahl M, Vestbo J, Zacho J, Lange P, Tybjærg-Hansen A, Nordestgaard BG. C reactive protein and chronic obstructive pulmonary disease: a Mendelian randomisation approach. Thorax 2011 Mar;66(3):197-204.
(113) Marott SC, Nordestgaard BG, Zacho J, et al. Does elevated C-reactive protein increase atrial fibrillation risk? A Mendelian randomization of 47,000 individuals from the general population. J Am Coll Cardiol 2010 Aug 31;56(10):789-95.
(114) Skaaby T, Husemoen LL, Martinussen T, et al. Vitamin D status, filaggrin genotype, and cardiovascular risk factors: a Mendelian randomization approach. PLoS One 2013;8(2):e57647.
(115) Cruchaga C, Kauwe JS, Nowotny P, et al. Cerebrospinal fluid APOE levels: an endophenotype for genetic studies for Alzheimer's disease. Hum Mol Genet 2012 Oct 15;21(20):4558-71.
(116) Hughes K, Flynn T, de Zoysa J, Dalbeth N, Merriman TR. Mendelian randomization analysis associates increased serum urate, due to genetic variation in uric acid transporters, with improved renal function. Kidney Int 2014 Feb;85(2):344-51.
(117) Benn M, Tybjærg-Hansen A, McCarthy MI, Jensen GB, Grande P, Nordestgaard BG. Nonfasting glucose, ischemic heart disease, and myocardial infarction: a Mendelian randomization study. J Am Coll Cardiol 2012 Jun 19;59(25):2356-65.
(118) Varbo A, Benn M, Tybjærg-Hansen A, Nordestgaard BG. Elevated remnant cholesterol causes both low-grade inflammation and ischemic heart disease, whereas elevated low-density lipoprotein cholesterol causes ischemic heart disease without inflammation. Circulation 2013 Sep 17;128(12):1298-309.
(119) Stender S, Nordestgaard BG, Tybjærg-Hansen A. Elevated body mass index as a causal risk factor for symptomatic gallstone disease: a Mendelian randomization study. Hepatology 2013 Dec;58(6):2133-41.
(120) Kamstrup PR, Nordestgaard BG. Lipoprotein(a) concentrations, isoform size, and risk of type 2 diabetes: a Mendelian randomisation study. The Lancet Diabetes & Endocrinology 2013 Nov;1(3):220-7.
(121) Wium-Andersen MK, Orsted DD, Nordestgaard BG. Elevated C-reactive protein associated with late- and very-late-onset schizophrenia in the general population: a prospective study. Schizophr Bull 2014 Sep;40(5):1117-27.
(122) Heikkilä K, Silander K, Salomaa V, et al. C-reactive protein-associated genetic variants and cancer risk: findings from FINRISK 1992, FINRISK 1997 and Health 2000 studies. Eur J Cancer 2011 Feb;47(3):404-12.
(123) Allin KH, Nordestgaard BG, Zacho J, Tybjærg-Hansen A, Bojesen SE. C-reactive protein and the risk of cancer: a mendelian randomization study. J Natl Cancer Inst 2010 Feb 3;102(3):202-6.
(124) Kamstrup PR, Tybjærg-Hansen A, Nordestgaard BG. Elevated lipoprotein(a) and risk of aortic valve stenosis in the general population. J Am Coll Cardiol 2014 Feb 11;63(5):470-7.

(125) Pierce BL, Tong L, Argos M, et al. Arsenic metabolism efficiency has a causal role in arsenic toxicity: Mendelian randomization and gene-environment interaction. International Journal of Epidemiology 2013 Dec;42(6):1862-71.
(126) Oelsner EC, Pottinger TD, Burkart KM, et al. Adhesion molecules, endothelin-1 and lung function in seven population-based cohorts. Biomarkers 2013 May;18(3):196-203.
(127) Trombetta M, Bonetti S, Boselli ML, et al. PPARG2 Pro12Ala and ADAMTS9 rs4607103 as "insulin resistance loci" and "insulin secretion loci" in Italian individuals. The GENFIEV study and the Verona Newly Diagnosed Type 2 Diabetes Study (VNDS) 4. Acta Diabetol 2013 Jun;50(3):401-8.
(128) Varbo A, Benn M, Tybjærg-Hansen A, Jørgensen AB, Frikke-Schmidt R, Nordestgaard BG. Remnant cholesterol as a causal risk factor for ischemic heart disease. J Am Coll Cardiol 2013 Jan 29;61(4):427-36.
(129) Jørgensen AB, Frikke-Schmidt R, West AS, Grande P, Nordestgaard BG, Tybjærg-Hansen A. Genetically elevated non-fasting triglycerides and calculated remnant cholesterol as causal risk factors for myocardial infarction. Eur Heart J 2013 Jun;34(24):1826-33.
(130) Haase CL, Tybjærg-Hansen A, Qayyum AA, Schou J, Nordestgaard BG, Frikke-Schmidt R. LCAT, HDL cholesterol and ischemic cardiovascular disease: a Mendelian randomization study of HDL cholesterol in 54,500 individuals. J Clin Endocrinol Metab 2012 Feb;97(2):E248-E256.
(131) Benn M, Tybjærg-Hansen A, Stender S, Frikke-Schmidt R, Nordestgaard BG. Low-density lipoprotein cholesterol and the risk of cancer: a mendelian randomization study. J Natl Cancer Inst 2011 Mar 16;103(6):508-19.
(132) Zacho J, Tybjærg-Hansen A, Nordestgaard BG. C-reactive protein and risk of venous thromboembolism in the general population. Arterioscler Thromb Vasc Biol 2010 Aug;30(8):1672-8.
(133) Wium-Andersen MK, Orsted DD, Nordestgaard BG. Elevated C-reactive protein, depression, somatic diseases, and all-cause mortality: a mendelian randomization study. Biol Psychiatry 2014 Aug 1;76(3):249-57.
(134) Stender S, Frikke-Schmidt R, Benn M, Nordestgaard BG, Tybjærg-Hansen A. Low-density lipoprotein cholesterol and risk of gallstone disease: a Mendelian randomization study and meta-analyses. J Hepatol 2013 Jan;58(1):126-33.
(135) Yaghootkar H, Lamina C, Scott RA, et al. Mendelian randomization studies do not support a causal role for reduced circulating adiponectin levels in insulin resistance and type 2 diabetes. Diabetes 2013 Oct;62(10):3589-98.
(136) Chen L, Davey Smith G, Harbord RM, Lewis SJ. Alcohol intake and blood pressure: a systematic review implementing a Mendelian randomization approach. PLoS Med 2008 Mar 4;5(3):e52.
(137) Shah S, Casas JP, Drenos F, et al. Causal relevance of blood lipid fractions in the development of carotid atherosclerosis: Mendelian randomization analysis. Circ Cardiovasc Genet 2013 Feb;6(1):63-72.
(138) Palmer TM, Nordestgaard BG, Benn M, et al. Association of plasma uric acid with ischaemic heart disease and blood pressure: mendelian randomisation analysis of two large cohorts. BMJ 2013;347:f4262.

(139) De Silva NM, Freathy RM, Palmer TM, et al. Mendelian randomization studies do not support a role for raised circulating triglyceride levels influencing type 2 diabetes, glucose levels, or insulin resistance. Diabetes 2011 Mar;60(3):1008-18.
(140) Thanassoulis G, Campbell CY, Owens DS, et al. Genetic associations with valvular calcification and aortic stenosis. N Engl J Med 2013 Feb 7;368(6):503-12.
(141) Nordestgaard BG, Palmer TM, Benn M, et al. The effect of elevated body mass index on ischemic heart disease risk: causal estimates from a Mendelian randomisation approach. PLoS Med 2012;9(5):e1001212.
(142) Elliott P, Chambers JC, Zhang W, et al. Genetic Loci associated with C-reactive protein levels and risk of coronary heart disease. JAMA 2009 Jul 1;302(1):37-48.
(143) Ye Z, Haycock PC, Gurdasani D, et al. The association between circulating lipoprotein(a) and type 2 diabetes: is it causal? Diabetes 2014 Jan;63(1):332-42.
(144) Voight BF, Peloso GM, Orho-Melander M, et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet 2012 Aug 11;380(9841):572-80.
(145) Mamasoula C, Prentice RR, Pierscionek T, et al. Association between C677T polymorphism of methylene tetrahydrofolate reductase and congenital heart disease: meta-analysis of 7697 cases and 13,125 controls. Circ Cardiovasc Genet 2013 Aug;6(4):347-53.
(146) Harrison SC, Smith AJ, Jones GT, et al. Interleukin-6 receptor pathways in abdominal aortic aneurysm. Eur Heart J 2013 Dec;34(48):3707-16.
(147) Stender S, Frikke-Schmidt R, Nordestgaard BG, Grande P, Tybjærg-Hansen A. Genetically elevated bilirubin and risk of ischaemic heart disease: three Mendelian randomization studies and a meta-analysis. J Intern Med 2013 Jan;273(1):59-68.
(148) Hingorani AD, Casas JP. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. Lancet 2012 Mar 31;379(9822):1214-24.
(149) Clarke R, Bennett DA, Parish S, et al. Homocysteine and coronary heart disease: meta-analysis of MTHFR case-control studies, avoiding publication bias. PLoS Med 2012 Feb;9(2):e1001177.
(150) Wang J, Wang H, Chen Y, Hao P, Zhang Y. Alcohol ingestion and colorectal neoplasia: a meta-analysis based on a Mendelian randomization approach. Colorectal Dis 2011 May;13(5):e71-e78.
(151) Casas JP, Ninio E, Panayiotou A, et al. PLA2G7 Genotype, lipoprotein-associated phospholipase A2 activity, and coronary heart disease risk in 10 494 cases and 15 624 controls of european ancestry. Circulation 2010 Jun 1;121(21):2284-93.
(152) Lewis SJ, Baker I, Davey Smith G. Meta-analysis of vitamin D receptor polymorphisms and pulmonary tuberculosis risk. International Journal of Tuberculosis and Lung Disease 2005 Oct;9(10):1174-7.
(153) Davies JR, Field S, Randerson-Moor J, et al. An inherited variant in the gene coding for vitamin D-binding protein and survival from cutaneous melanoma: a BioGenoMEL study. Pigment Cell Melanoma Res 2014 Mar;27(2):234-43.
(154) Panoutsopoulou K, Metrustry S, Doherty SA, et al. The effect of FTO variation on increased osteoarthritis risk is mediated through body mass index: a mendelian randomisation study. Ann Rheum Dis 2013 Aug 6.

(155) Ioannidis A, Ikonomi E, Dimou NL, Douma L, Bagos PG. Polymorphisms of the insulin receptor and the insulin receptor substrates genes in polycystic ovary syndrome: a Mendelian randomization meta-analysis. Mol Genet Metab 2010 Feb;99(2):174-83.
(156) Rice NE, Bandinelli S, Corsi AM, et al. The paraoxonase (PON1) Q192R polymorphism is not associated with poor health status or depression in the ELSA or INCHIANTI studies. Int J Epidemiol 2009 Oct;38(5):1374-9.
(157) Rafiq S, Melzer D, Weedon MN, et al. Gene variants influencing measures of inflammation or predisposing to autoimmune and inflammatory diseases are not associated with the risk of type 2 diabetes. Diabetologia 2008 Dec;51(12):2205-13.
(158) Pfister R, Sharp S, Luben R, et al. Mendelian randomization study of B-type natriuretic peptide and type 2 diabetes: evidence of causal association from population studies. PLoS Med 2011 Oct;8(10):e1001112.
(159) Pfister R, Barnes D, Luben R, et al. No evidence for a causal link between uric acid and type 2 diabetes: a Mendelian randomisation approach. Diabetologia 2011 Oct;54(10):2561-9.
(160) Marjot T, Yadav S, Hasan N, Bentley P, Sharma P. Genes associated with adult cerebral venous thrombosis. Stroke 2011 Apr;42(4):913-8.
(161) Sarwar N, Sandhu MS, Ricketts SL, et al. Triglyceride-mediated pathways and coronary disease: collaborative analysis of 101 studies. Lancet 2010 May 8;375(9726):1634-9.
(162) Bentley P, Peck G, Smeeth L, Whittaker J, Sharma P. Causal relationship of susceptibility genes to ischemic stroke: comparison to ischemic heart disease and biochemical determinants. PLoS One 2010;5(2):e9136.
(163) Perry JR, Ferrucci L, Bandinelli S, et al. Circulating beta-carotene levels and type 2 diabetes-cause or effect? Diabetologia 2009 Oct;52(10):2117-21.
(164) Perry JR, Weedon MN, Langenberg C, et al. Genetic evidence that raised sex hormone binding globulin (SHBG) levels reduce the risk of type 2 diabetes. Hum Mol Genet 2010 Feb 1;19(3):535-44.
(165) Boccia S, Hashibe M, Galli P, et al. Aldehyde dehydrogenase 2 and head and neck cancer: a meta-analysis implementing a Mendelian randomization approach. Cancer Epidemiol Biomarkers Prev 2009 Jan;18(1):248-54.
(166) Casas JP, Bautista LE, Smeeth L, Sharma P, Hingorani AD. Homocysteine and stroke: evidence on a causal link from mendelian randomisation. Lancet 2005 Jan 15;365(9455):224-32.
(167) Casas JP, Shah T, Cooper J, et al. Insight into the nature of the CRP-coronary event association using Mendelian randomization. Int J Epidemiol 2006 Aug;35(4):922-31.
(168) Fall T, Hagg S, Magi R, et al. The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis. PLoS Med 2013;10(6):e1001474.
(169) Pichler I, Del Greco MF, Gögele M, et al. Serum iron levels and the risk of Parkinson disease: a Mendelian randomization study. PLoS Med 2013;10(6):e1001462.
(170) Vimaleswaran KS, Berry DJ, Lu C, et al. Causal relationship between obesity and vitamin D status: bi-directional Mendelian randomization analysis of multiple cohorts. PLoS Med 2013;10(2):e1001383.
(171) Ference BA, Yoo W, Alesh I, et al. Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: a Mendelian randomization analysis. J Am Coll Cardiol 2012 Dec 25;60(25):2631-9.

(172) Niu W, Liu Y, Qi Y, Wu Z, Zhu D, Jin W. Association of interleukin-6 circulating levels with coronary artery disease: a meta-analysis implementing mendelian randomization approach. Int J Cardiol 2012 May 31;157(2):243-52.
(173) Niu W, Zhang X, Qi Y. Association of an apolipoprotein E polymorphism with circulating cholesterols and hypertension: a meta-based Mendelian randomization analysis. Hypertens Res 2012 Apr;35(4):434-40.
(174) Huang T, Ren J, Huang J, Li D. Association of homocysteine with type 2 diabetes: a meta-analysis implementing Mendelian randomization approach. BMC Genomics 2013;14:867.
(175) Davey Smith G, Harbord R, Milton J, Ebrahim S, Sterne JAC. Does elevated plasma fibrinogen increase the risk of coronary heart disease? Evidence from a meta-analysis of genetic association studies. Arteriosclerosis, Thrombosis, and Vascular Biology 2005 Oct;25(10):2228-33.
(176) Lawlor DA, Harbord RM, Timpson NJ, et al. The association of C-reactive protein and CRP genotype with coronary heart disease: findings from five studies with 4,610 cases amongst 18,637 participants. PLoS One 2008;3(8):e3011.
(177) Kamstrup PR, Tybjærg-Hansen A, Steffensen R, Nordestgaard BG. Genetically elevated lipoprotein(a) and increased risk of myocardial infarction. JAMA 2009 Jun 10;301(22):2331-9.
(178) Kamstrup PR, Tybjærg-Hansen A, Nordestgaard BG. Genetic evidence that lipoprotein(a) associates with atherosclerotic stenosis rather than venous thrombosis. Arterioscler Thromb Vasc Biol 2012 Jul;32(7):1732-41.
(179) Stegeman BH, Helmerhorst FM, Vos HL, Rosendaal FR, Van Hylckama Vlieg A. Sex hormone-binding globulin levels are not causally related to venous thrombosis risk in women not using hormonal contraceptives. Journal of Thrombosis and Haemostasis 2012 Oct;10(10):2061-7.
(180) Adamsson Eryd S, Sjögren M, Smith JG, et al. Ceruloplasmin and atrial fibrillation: evidence of causality from a population-based Mendelian randomization study. J Intern Med 2014 Feb;275(2):164-71.
(181) Ding EL, Song Y, Manson JE, et al. Sex hormone-binding globulin and risk of type 2 diabetes in women and men. N Engl J Med 2009 Sep 17;361(12):1152-63.
(182) Husemoen LL, Skaaby T, Martinussen T, et al. Investigating the causal effect of vitamin D on serum adiponectin using a mendelian randomization approach. Eur J Clin Nutr 2014 Feb;68(2):189-95.
(183) Song Y, Yeung E, Liu A, et al. Pancreatic beta-cell function and type 2 diabetes risk: quantify the causal effect using a Mendelian randomization approach based on meta-analyses. Hum Mol Genet 2012 Nov 15;21(22):5010-8.
(184) Wensley F, Gao P, Burgess S, et al. Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. BMJ 2011;342:d548.
(185) Zhao J, Jiang C, Lam TH, et al. Genetically predicted testosterone and cardiovascular risk factors in men: a Mendelian randomization analysis in the Guangzhou Biobank Cohort Study. Int J Epidemiol 2014 Feb;43(1):140-8.
(186) Holmes MV, Simon T, Exeter HJ, et al. Secretory phospholipase A(2)-IIA and cardiovascular disease: a mendelian randomization study. J Am Coll Cardiol 2013 Nov 19;62(21):1966-76.

(187) Linsel-Nitschke P, Götz A, Erdmann J, et al. Lifelong reduction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of coronary artery disease--a Mendelian Randomisation study. PLoS One 2008;3(8):e2986.
(188) Kunutsor SK, Burgess S, Munroe PB, Khan H. Vitamin D and high blood pressure: causal association or epiphenomenon? Eur J Epidemiol 2014 Jan;29(1):1-14.
(189) Bochud M, Rousson V. Usefulness of Mendelian randomization in observational epidemiology. Int J Environ Res Public Health 2010 Mar;7(3):711-28.
(190) VanderWeele TJ, Tchetgen Tchetgen EJ, Cornelis M, Kraft P. Methodological challenges in mendelian randomization. Epidemiology 2014 May;25(3):427-35.
(191) Boef AG, Dekkers OM, le Cessie S, Vandenbroucke JP. Reporting instrumental variable analyses. Epidemiology 2013 Nov;24(6):937-8.
(192) Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet 2014 Sep 15;23(R1):R89-R98.
(193) Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization. Int J Epidemiol 2013 Aug;42(4):1134-44.
(194) Palmer TM, Sterne JA, Harbord RM, et al. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. Am J Epidemiol 2011 Jun 15;173(12):1392-403.
(195) Rassen JA, Schneeweiss S, Glynn RJ, Mittleman MA, Brookhart MA. Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. Am J Epidemiol 2009 Feb 1;169(3):273-84.
(196) Vuistiner P, Bochud M, Rousson V. A comparison of three methods of Mendelian randomization when the genetic instrument, the risk factor and the outcome are all binary. PLoS One 2012;7(5):e35951.
(197) Burgess S. Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. Stat Med 2013 Nov 30;32(27):4726-47.
(198) Clarke PS, Windmeijer F. Instrumental Variable Estimators for Binary Outcomes. Journal of the American Statistical Association 2012 Oct 8;107(500):1638-52.

Box 1 Proposed checklist for reporting Mendelian randomisation studies

Methods

- If an expected genotype-outcome association is calculated, report how this was calculated and how the confidence interval was obtained. Take into account the variance of both the genotype-exposure and the exposure-outcome association.
- If an instrumental variable analysis is performed, report in detail which method was used and how the confidence interval was obtained. For non-standard instrumental variable methods (e.g. methods used to estimate an odds ratio), discuss the properties of these methods.
- If data from multiple populations are used, clearly explain how and at what stage the data/estimates were combined.

Results

- Report the strength of the association between the genetic instrument and the exposure, using a partial F-statistic if possible.
- Show the association between the genetic instrument and measured confounders. If multiple study populations are used, show this for those populations for which this information is available.
- Report the association of the genotype and the outcome

Discussion

- Discuss the plausibility of the second and third instrumental variable assumptions in the specific setting of the study: could pleiotropy, linkage disequilibrium, canalisation, population stratification or unmeasured confounding of the genotype-outcome relation affect results in this study?

Supplementary methods: Search strategy

PubMed
Restrictions: date 01/01/2003-31/12/2013
Query:
("Mendelian Randomization Analysis"[Mesh] OR "Mendelian randomisation"[all fields] OR "Mendelian randomization"[all fields] OR (Mendelian[all fields] AND randomi*[all fields]) OR "genetic instrumental variable"[all fields] OR "genetic instrumental variables"[all fields] OR "genetic instrument"[all fields] OR "genetic instruments"[all fields] OR "genes as instruments"[all fields] OR "gene as instrument"[all fields] OR "genes as instrument"[all fields] OR "gene as instruments"[all fields] OR (instrument*[ti] AND (gene[ti] OR genes[ti] OR genetic*[ti] OR mendel*[ti])) OR (("instrumental variable"[all fields] OR "instrumental variables"[all fields] OR "instrumented analysis"[all fields] OR "instrumented analyses"[all fields] OR "instrumental variable analysis"[all fields] OR "instrumental variable analyses"[all fields] OR "instrumental variables analysis"[all fields] OR "instrumental variables analyses"[all fields]) **AND** (gene OR genes OR genetics OR mendel OR mendelian)) OR ("mendelian"[all fields] AND ("randomisation"[all fields] OR "randomization"[all fields] OR "randomising"[all fields] OR "randomizing"[all fields])))

Medline
Restrictions: year 2003-2013
Query:
(Mendelian Randomization Analysis/ OR "Mendelian randomisation".af OR "Mendelian randomization".af OR "genetic instrumental variable".af OR "genetic instrumental variables".af OR "genetic instrument".af OR "genetic instruments".af OR "mendel randomise" OR "mendel randomize" OR "mendel randomization" OR "mendel randomisation" OR "random Mendelian" OR "genes as instruments".af OR "gene as instrument".af OR "genes as instrument".af OR "gene as instruments".af) OR (instrument*.ti AND (gene.ti OR genes.ti OR genetic*.ti OR mendel*.ti)) OR (("instrumental variable".af OR "instrumental variables".af OR "instrumented analysis".af OR "instrumented analyses".af OR "instrumental variable analysis".af OR "instrumental variable analyses".af OR "instrumental variables analysis".af OR "instrumental variables analyses".af) **AND** (gene OR genes OR genetics OR mendel OR mendelian).af) OR ("mendelian".af AND ("randomisation".af OR "randomization".af OR "randomising".af OR "randomizing".af))

Embase

Restrictions: year 2003-2013, no conference abstracts.

Query:

(Mendelian Randomization Analysis/ OR "Mendelian randomisation".af OR "Mendelian randomization".af OR "genetic instrumental variable".af OR "genetic instrumental variables".af OR "genetic instrument".af OR "genetic instruments". af OR "mendel randomise" OR "mendel randomize" OR "mendel randomization" OR "mendel randomisation" OR "random Mendelian" OR "genes as instruments". af OR "gene as instrument".af OR "genes as instrument".af OR "gene as instruments". af) OR (instrument*.ti AND (gene.ti OR genes.ti OR genetic*.ti OR mendel*.ti)) OR (("instrumental variable".af OR "instrumental variables".af OR "instrumented analysis".af OR "instrumented analyses".af OR "instrumental variable analysis".af OR "instrumental variable analyses".af OR "instrumental variables analysis".af OR "instrumental variables analyses".af) **AND** (gene OR genes OR genetics OR mendel OR mendelian).af) OR ("mendelian".af AND ("randomisation".af OR "randomization". af OR "randomising".af OR "randomizing".af))

Web of Science

Restrictions: year 2003-2013, no conference abstracts.

Query:

TI=(("Mendelian randomisation" OR "Mendelian randomization" OR "genetic instrumental variable" OR "genetic instrumental variables" OR "genetic instrument" OR "genetic instruments" OR "mendel randomise" OR "mendel randomize" OR "mendel randomization" OR "mendel randomisation" OR "random Mendelian" OR "genes as instruments" OR "gene as instrument" OR "genes as instrument" OR "gene as instruments" OR "instrumental genetic variable" OR "instrumental genetic variable")

Supplementary Table 1 Exposures and genetic instruments used .

Exposure	Number of studies	Genes in which variation was used as an instrument
C-reactive protein	29	*CRP*,[1-28] *LEPR*,[13;14] *HNF1A*,[7;14] *IL6R*,[7;14;24] *APOE*[7;14], *genetic risk score*[29]
BMI/ fat mass/ percentage body fat	25	*FTO*,[8;12;30-45] *MC4R*,[12;32;37-39;41;43;44] *TMEM18*,[32;37;38;44] *VEGFA*, [46]*genetic risk score*[47-52]
Alcohol use	12	*ALDH2*,[53-58] *ADH1B*,[59-63] *ADH1C*[61;64]
Vitamin D levels	10	*GC*,[65-69] *DHCR7/NADSYN1*,[65-67;70] *CYP2R1*,[65-67;70] *CYP24A1*,[66;67] *FLG*,[68;71] *VDR*,[72] *genetic risk scores*[47;73]
Homocysteïne	8	*MTHFR*[74-81]
"Folate metabolism"	4	*MTHFR*[82-85]
LDL-cholesterol	8	*SORT1*,[86] *PCSK9*,[86-89] *LDLR*,[86-88;90] *HMGCR*,[86] *ABCG8*,[86;89] *APOE*,[86;88;89;91] *APOB*,[87;88] *genetic risk score*[24;92]
HDL-cholesterol	5	*LIPC*,[87;93] *LIPG*[94], *ABCA1*,[87] *LCAT*,[95] *genetic risk score*[92;94]
Total cholesterol	4	*APOE*[91;96-98]
Remnant cholesterol	3	*APOA5*,[87;99] *TRIB1*,[87] *GCKR*[87], *genetic risk score*[24]
Remnant cholesterol:HDL ratio	1	*LPL*[87]
Triglycerides	5	*APOA5*,[99-101] *genetic risk score*[92;102]
Lipoprotein(a)	7	*LPA*[103-109]
Lp-PLA_2 (activity)	4	*PLA2G7*,[15;110;111] *PLA2G2A*[112]
ApoAI	1	*APOA5-A4-C3-A1*[15]
ApoB	1	*APOB*[15]
Uric acid	7	*SLC2A9*,[32;37;49;113-115] *ABCG2*,[113;115] *SLC17A1*,[115] *SLC22A11*,[115] *SLC22A12*,[115] *genetic risk score*[116]
IL-6/ IL-6 receptor signalling	5	*IL6*,[117;118] *IL6R*[26;119;120]
Fetuin-A	4	*AHSG*[35;121-123]
Adiponectin	4	*ADIPOQ*[124-127]
Fibrinogen	4	*FIBA-B-G cluster*,[15] *FGB*[128-130]
Fasting glucose	3	*genetic risk score*[50;131;132]
HOMA-IR	2	*GCKR*,[125] *ADAMTS9*,[133] *PPARG2*[133]
Beta-cell function	2	*ADAMTS9*,[133] *TCF7L2*[134]
Non-fasting glucose	1	*GCK*,[135] *G6PC2*,[135] *ADCY5*,[135] *DGKB*,[135] *ADRA2A*[135]
Fasting insulin	1	*INSR*,[136] *IRS1*[136]
Type 2 diabetes	2	*genetic risk score*[131;137]
Type 1 diabetes	1	*genetic risk score*[137]
Milk consumption	3	*LCT*[138-140]
Iron status (ferritin/serum iron)	3	*HFE*,[141;142] *TMPRSS6*[141;143]
Bilirubin	3	*UGT1A1*[144-146]
SHBG	3	*SHBG*[147-149]
Testosterone	2	*SHBG*,[150] *FAM9B*,[150] *CYP19A1*,[151] *ESR2*[151]

Prenatal testosterone exposure	1	*Sex of co-twin*[152]
Caffeine (intake)	2	*CYP1A2,*[153;154] *NAT2,*[154] *GSTA1*[154]
Vitamin B-12	2	*FUT2,*[155;156] *TCN2,*[155] *CUBN*[156]
Total transcobalamin	1	*TCN2*[156]
Smoking	2	*CHRNA5–CHRNA3–CHRNB4 cluster*[157;158]
PAI-1 levels	2	*PAI14G/5G*[26;78]
Malaria infection	2	HbAS phenotype[159;160]
IL-18	2	*IL18*[26;161]
Macrophage migration inhibitory factor	2	*MIF*[26;162]
6-propylthyouracil tasting	1	*TAS2R38*[163]
Monocyte chemotactic protein-1	1	*CCL2*[1]
Leukocyte telomere length	1	*genetic risk score*[164]
Triacylglycerol	1	*genetic risk score*[50]
$sPLA_2$-IIa	1	*PLA2G2A*[165]
γ-glutamyltransferase	1	*GGT1*[166]
Δ^5-desaturase and Δ^6-desaturase activity	1	*FADS1*[167]
Monocyte CD36 expression	1	*CD36*[168]
Factor VII	1	*F7*[15]
Retinol-binding protein 4	1	*RBP4*[169]
Complement factor H	1	*CFH*[170]
Surfactant protein D	1	*SP-D*[171]
MiR-34b	1	*Pri-miR-34b/c*[172]
ICAM-1	1	*ICAM1*[173]
P-selectin	1	*SELP*[173]
CSF ApoE	1	*APOE,*[174] *genetic risk score*[174]
NT-pro-BNP	1	*BNP*[175]
APC resistance	1	*FVL*[78]
ACE activity	1	*ACE D/I*[78]
Prothrombin levels	1	*F2*[78]
Beta-carotene	1	*BCMO1*[176]
Arsenic metabolism efficiency	1	*AS3MT*[177]
IL-1RA	1	*IL1RN*[26]
Inflammatory/auto-immune disease	1	*IL23R, PTPN2, PTPN22, SH2B3, IL2RA (+ 31 others)*[26]
Ceruloplasmin	1	*CP*[178]
Organophosphate exposure	1	*PON1*[179]

Supplementary Table 2 *Type of genetic instrument in the instrumental variable studies*

Type of genetic instrument	1 Study population	Refs	>1 Study population*	Refs
Single SNP/allele	27	8; 19; 23; 31-36; 46; 49; 53; 54; 74; 93; 105; 114; 122; 123; 125; 142; 146; 158; 160; 161; 163; 166	16	16; 24; 30; 37; 57; 81; 91; 94; 95; 103; 109; 112; 117; 134; 136; 175
Multiple SNPs in separate analyses	6	13; 39; 66; 126; 153; 156	12	14; 68; 86; 87; 89; 107; 108; 133; 135; 141; 148; 174
Multiple SNPs in a single analysis	11	13; 39; 41; 43; 61; 65; 83; 121; 150; 156; 164	14	37; 68; 71; 87-89; 108; 124; 133; 135; 148; 151; 173; 177
Combinations of SNPs, e.g. haplotypes	6	10; 17-20; 22	10	3; 5; 6; 9; 11; 24; 27; 28; 99; 130
Genetic risk score/allele score etc.	8	32; 48-51; 66; 105; 126; 164	11	7; 24; 38; 44; 47; 87; 92; 94; 102; 115; 174
Repeats	0	-	4	104; 106-108

Abbreviations: SNP single nucleotide polymorphism.
The total number of types of genetic instrument exceeds the total number of studies (49 in 1 study population, 52 in >1 study population) because some studies used multiple types of genetic instrument

References

(1) Alegret JM, Aragonès G, Elosua R et al. The relevance of the association between inflammation and atrial fibrillation. *Eur J Clin Invest* 2013;43:324-331.

(2) Rius-Ottenheim N, de Craen AJ, Geleijnse JM et al. C-reactive protein haplotypes and dispositional optimism in obese and nonobese elderly subjects. *Inflamm Res* 2012;61:43-51.

(3) Wensley F, Gao P, Burgess S et al. Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ* 2011;342:d548.

(4) Bolton CE, Schumacher W, Cockcroft JR et al. The CRP genotype, serum levels and lung function in men: the Caerphilly Prospective Study. *Clin Sci (Lond)* 2011;120:347-355.

(5) Dahl M, Vestbo J, Zacho J, Lange P, Tybjærg-Hansen A, Nordestgaard BG. C reactive protein and chronic obstructive pulmonary disease: a Mendelian randomisation approach. *Thorax* 2011;66:197-204.

(6) Marott SC, Nordestgaard BG, Zacho J et al. Does elevated C-reactive protein increase atrial fibrillation risk? A Mendelian randomization of 47,000 individuals from the general population. *J Am Coll Cardiol* 2010;56:789-795.

(7) Heikkilä K, Silander K, Salomaa V et al. C-reactive protein-associated genetic variants and cancer risk: findings from FINRISK 1992, FINRISK 1997 and Health 2000 studies. *Eur J Cancer* 2011;47:404-412.

(8) Timpson NJ, Nordestgaard BG, Harbord RM et al. C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. *Int J Obes (Lond)* 2011;35:300-308.

(9) Zacho J, Tybjærg-Hansen A, Nordestgaard BG. C-reactive protein and risk of venous thromboembolism in the general population. *Arterioscler Thromb Vasc Biol* 2010;30:1672-1678.

(10) Zacho J, Tybjærg-Hansen A, Nordestgaard BG. C-reactive protein and all-cause mortality--the Copenhagen City Heart Study. *Eur Heart J* 2010;31:1624-1632.

(11) Allin KH, Nordestgaard BG, Zacho J, Tybjærg-Hansen A, Bojesen SE. C-reactive protein and the risk of cancer: a mendelian randomization study. *J Natl Cancer Inst* 2010;102:202-206.

(12) Welsh P, Polisecki E, Robertson M et al. Unraveling the directional link between adiposity and inflammation: a bidirectional Mendelian randomization approach. *J Clin Endocrinol Metab* 2010;95:93-99.

(13) Bochud M, Marquant F, Marques-Vidal PM et al. Association between C-reactive protein and adiposity in women. *J Clin Endocrinol Metab* 2009;94:3969-3977.

(14) Elliott P, Chambers JC, Zhang W et al. Genetic Loci associated with C-reactive protein levels and risk of coronary heart disease. *JAMA* 2009;302:37-48.

(15) Drenos F, Talmud PJ, Casas JP et al. Integrated associations of genotypes with multiple blood biomarkers linked to coronary heart disease risk. *Hum Mol Genet* 2009;18:2305-2316.

(16) Lawlor DA, Harbord RM, Timpson NJ et al. The association of C-reactive protein and CRP genotype with coronary heart disease: findings from five studies with 4,610 cases amongst 18,637 participants. *PLoS One* 2008;3:e3011.

(17) Kivimäki M, Lawlor DA, Davey Smith G et al. Does high C-reactive protein concentration increase atherosclerosis? The Whitehall II Study. *PLoS One* 2008;3:e3013.
(18) Brunner EJ, Kivimäki M, Witte DR et al. Inflammation, insulin resistance, and diabetes--Mendelian randomization using CRP haplotypes points upstream. *PLoS Med* 2008;5:e155.
(19) Sunyer J, Pistelli R, Plana E et al. Systemic inflammation, genetic susceptibility and lung function. *Eur Respir J* 2008;32:92-97.
(20) Viikari LA, Huupponen RK, Viikari JS et al. Relationship between leptin and C-reactive protein in young Finnish adults. *J Clin Endocrinol Metab* 2007;92:4753-4758.
(21) Casas JP, Shah T, Cooper J et al. Insight into the nature of the CRP-coronary event association using Mendelian randomization. *Int J Epidemiol* 2006;35:922-931.
(22) Timpson NJ, Lawlor DA, Harbord RM et al. C-reactive protein and its role in metabolic syndrome: mendelian randomisation study. *Lancet* 2005;366:1954-1959.
(23) Davey Smith G, Lawlor DA, Harbord R et al. Association of C-reactive protein with blood pressure and hypertension: life course confounding and mendelian randomization tests of causality. *Arterioscler Thromb Vasc Biol* 2005;25:1051-1056.
(24) Varbo A, Benn M, Tybjærg-Hansen A, Nordestgaard BG. Elevated remnant cholesterol causes both low-grade inflammation and ischemic heart disease, whereas elevated low-density lipoprotein cholesterol causes ischemic heart disease without inflammation. *Circulation* 2013;128:1298-1309.
(25) Veen G, Giltay EJ, Van Vliet IM et al. C-reactive protein polymorphisms are associated with the cortisol awakening response in basal conditions in human subjects. *Stress* 2011;14:128-135.
(26) Rafiq S, Melzer D, Weedon MN et al. Gene variants influencing measures of inflammation or predisposing to autoimmune and inflammatory diseases are not associated with the risk of type 2 diabetes. *Diabetologia* 2008;51:2205-2213.
(27) Wium-Andersen MK, Orsted DD, Nordestgaard BG. Elevated C-reactive protein, depression, somatic diseases, and all-cause mortality: a mendelian randomization study. *Biol Psychiatry* 2014;76:249-257.
(28) Wium-Andersen MK, Orsted DD, Nordestgaard BG. Elevated C-reactive protein associated with late- and very-late-onset schizophrenia in the general population: a prospective study. *Schizophr Bull* 2014;40:1117-1127.
(29) Oei L, Campos-Obando N, Dehghan A et al. Dissecting the relationship between high-sensitivity serum C-reactive protein and increased fracture risk: the Rotterdam Study. *Osteoporos Int* 2014;25:1247-1254.
(30) Fall T, Hagg S, Magi R et al. The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis. *PLoS Med* 2013;10:e1001474.
(31) Warodomwichit D, Sritara C, Thakkinstian A et al. Causal inference of the effect of adiposity on bone mineral density in adults. *Clin Endocrinol (Oxf)* 2013;78:694-699.
(32) Lyngdoh T, Vuistiner P, Marques-Vidal P et al. Serum uric acid and adiposity: deciphering causality using a bidirectional Mendelian randomization approach. *PLoS One* 2012;7:e39321.
(33) Kivimäki M, Jokela M, Hamer M et al. Examining overweight and obesity as risk factors for common mental disorders using fat mass and obesity-associated (FTO) genotype-instrumented analysis: The Whitehall II Study, 1985-2004. *Am J Epidemiol* 2011;173:421-429.

(34) Lawlor DA, Timpson NJ, Harbord RM et al. Exploring the developmental overnutrition hypothesis using parental-offspring associations and FTO as an instrumental variable. *PLoS Med* 2008;5:e33.
(35) Thakkinstian A, Chailurkit L, Warodomwichit D et al. Causal relationship between body mass index and fetuin-A level in the asian population: a bidirectional mendelian randomization study. *Clin Endocrinol (Oxf)* 2014;81:197-203.
(36) Kivimäki M, Davey Smith G, Timpson NJ et al. Lifetime body mass index and later atherosclerosis risk in young adults: examining causal links using Mendelian randomization in the Cardiovascular Risk in Young Finns study. *Eur Heart J* 2008;29:2552-2560.
(37) Palmer TM, Nordestgaard BG, Benn M et al. Association of plasma uric acid with ischaemic heart disease and blood pressure: mendelian randomisation analysis of two large cohorts. *BMJ* 2013;347:f4262.
(38) Nordestgaard BG, Palmer TM, Benn M et al. The effect of elevated body mass index on ischemic heart disease risk: causal estimates from a Mendelian randomisation approach. *PLoS Med* 2012;9:e1001212.
(39) Lawlor DA, Harbord RM, Tybjærg-Hansen A et al. Using genetic loci to understand the relationship between adiposity and psychological distress: a Mendelian Randomization study in the Copenhagen General Population Study of 53,221 adults. *J Intern Med* 2011;269:525-537.
(40) Brennan P, McKay J, Moore L et al. Obesity and cancer: Mendelian randomization approach utilizing the FTO genotype. *Int J Epidemiol* 2009;38:971-975.
(41) Timpson NJ, Harbord R, Davey Smith G, Zacho J, Tybjærg-Hansen A, Nordestgaard BG. Does greater adiposity increase blood pressure and hypertension risk?: Mendelian randomization using the FTO/MC4R genotype. *Hypertension* 2009;54:84-90.
(42) Lim LS, Tai ES, Aung T et al. Relation of age-related cataract with obesity and obesity genes in an Asian population. *Am J Epidemiol* 2009;169:1267-1274.
(43) Timpson NJ, Sayers A, Davey Smith G, Tobias JH. How does body fat influence bone mass in childhood? A Mendelian randomization approach. *J Bone Miner Res* 2009;24:522-533.
(44) Stender S, Nordestgaard BG, Tybjærg-Hansen A. Elevated body mass index as a causal risk factor for symptomatic gallstone disease: a Mendelian randomization study. *Hepatology* 2013;58:2133-2141.
(45) Panoutsopoulou K, Metrustry S, Doherty SA et al. The effect of FTO variation on increased osteoarthritis risk is mediated through body mass index: a mendelian randomisation study. *Ann Rheum Dis* 2013.
(46) Debette S, Wolf C, Lambert JC et al. Abdominal obesity and lower gray matter volume: a Mendelian randomization study. *Neurobiol Aging* 2014;35:378-386.
(47) Vimaleswaran KS, Berry DJ, Lu C et al. Causal relationship between obesity and vitamin D status: bi-directional Mendelian randomization analysis of multiple cohorts. *PLoS Med* 2013;10:e1001383.
(48) Jokela M, Elovainio M, Keltikangas-Järvinen L et al. Body mass index and depressive symptoms: instrumental-variables regression with genetic risk score. *Genes Brain Behav* 2012.

(49) Oikonen M, Wendelin-Saarenhovi M, Lyytikäinen LP et al. Associations between serum uric acid and markers of subclinical atherosclerosis in young adults. The cardiovascular risk in Young Finns study. *Atherosclerosis* 2012;223:497-503.
(50) Islam M, Jafar TH, Wood AR et al. Multiple genetic variants explain measurable variance in type 2 diabetes-related traits in Pakistanis. *Diabetologia* 2012;55:2193-2204.
(51) Mumby HS, Elks CE, Li S et al. Mendelian Randomisation Study of Childhood BMI and Early Menarche. *J Obes* 2011;2011:180729.
(52) Di Paola R, Marucci A, Fontana A et al. Role of obesity on all-cause mortality in whites with type 2 diabetes from Italy. *Acta Diabetologica* 2013;50:971-976.
(53) Au Yeung SL, Jiang C, Cheng KK et al. Moderate alcohol use and cardiovascular disease from Mendelian randomization. *PLoS One* 2013;8:e68054.
(54) Au Yeung SL, Jiang CQ, Cheng KK et al. Evaluation of moderate alcohol use and cognitive function among men using a Mendelian randomization design in the Guangzhou biobank cohort study. *Am J Epidemiol* 2012;175:1021-1028.
(55) Wang J, Wang H, Chen Y, Hao P, Zhang Y. Alcohol ingestion and colorectal neoplasia: a meta-analysis based on a Mendelian randomization approach. *Colorectal Dis* 2011;13:e71-e78.
(56) Boccia S, Hashibe M, Galli P et al. Aldehyde dehydrogenase 2 and head and neck cancer: a meta-analysis implementing a Mendelian randomization approach. *Cancer Epidemiol Biomarkers Prev* 2009;18:248-254.
(57) Chen L, Davey Smith G, Harbord RM, Lewis SJ. Alcohol intake and blood pressure: a systematic review implementing a Mendelian randomization approach. *PLoS Med* 2008;5:e52.
(58) Irons DE, McGue M, Iacono WG, Oetting WS. Mendelian randomization: a novel test of the gateway hypothesis and models of gene-environment interplay. *Dev Psychopathol* 2007;19:1181-1195.
(59) Humphriss R, Hall A, May M, Zuccolo L, Macleod J. Prenatal alcohol exposure and childhood balance ability: findings from a UK birth cohort study. *BMJ Open* 2013;3.
(60) Shaheen SO, Rutterford C, Zuccolo L et al. Prenatal alcohol exposure and childhood atopic disease: a Mendelian randomization approach. *J Allergy Clin Immunol* 2014;133:225-232.
(61) Lawlor DA, Nordestgaard BG, Benn M, Zuccolo L, Tybjærg-Hansen A, Davey Smith G. Exploring causal associations between alcohol and coronary heart disease risk factors: findings from a Mendelian randomization study in the Copenhagen General Population Study. *Eur Heart J* 2013;34:2519-2528.
(62) Zuccolo L, Lewis SJ, Davey Smith G et al. Prenatal alcohol exposure and offspring cognition and school performance. A 'Mendelian randomization' natural experiment. *International Journal of Epidemiology* 2013;42:1358-1370.
(63) Almeida OP, Hankey GJ, Yeap BB, Golledge J, Flicker L. The triangular association of ADH1B genetic polymorphism, alcohol consumption and the risk of depression in older men. *Mol Psychiatry* 2013.
(64) Heidrich J, Wellmann J, Doring A, Illig T, Keil U. Alcohol consumption, alcohol dehydrogenase and risk of coronary heart disease in the MONICA/KORA-Augsburg cohort 1994/1995-2002. *European Journal of Cardiovascular Prevention and Rehabilitation* 2007;14:769-774.

(65) Trummer O, Pilz S, Hoffmann MM et al. Vitamin D and mortality: a Mendelian randomization study. *Clin Chem* 2013;59:793-797.
(66) Theodoratou E, Palmer T, Zgaga L et al. Instrumental variable estimation of the causal effect of plasma 25-hydroxy-vitamin D on colorectal cancer risk: a mendelian randomization analysis. *PLoS One* 2012;7:e37662.
(67) Kunutsor SK, Burgess S, Munroe PB, Khan H. Vitamin D and high blood pressure: causal association or epiphenomenon? *Eur J Epidemiol* 2014;29:1-14.
(68) Husemoen LL, Skaaby T, Martinussen T et al. Investigating the causal effect of vitamin D on serum adiponectin using a mendelian randomization approach. *Eur J Clin Nutr* 2014;68:189-195.
(69) Davies JR, Field S, Randerson-Moor J et al. An inherited variant in the gene coding for vitamin D-binding protein and survival from cutaneous melanoma: a BioGenoMEL study. *Pigment Cell Melanoma Res* 2014;27:234-243.
(70) Yarwood A, Martin P, Bowes J et al. Enrichment of vitamin D response elements in RA-associated loci supports a role for vitamin D in the pathogenesis of RA. *Genes and Immunity* 2013;14:325-329.
(71) Skaaby T, Husemoen LL, Martinussen T et al. Vitamin D status, filaggrin genotype, and cardiovascular risk factors: a Mendelian randomization approach. *PLoS One* 2013;8:e57647.
(72) Lewis SJ, Baker I, Davey Smith G. Meta-analysis of vitamin D receptor polymorphisms and pulmonary tuberculosis risk. *International Journal of Tuberculosis and Lung Disease* 2005;9:1174-1177.
(73) Bonilla C, Gilbert R, Kemp JP et al. Using genetic proxies for lifecourse sun exposure to assess the causal relationship of sun exposure with circulating vitamin d and prostate cancer risk. *Cancer Epidemiol Biomarkers Prev* 2013;22:597-606.
(74) Lee HA, Park EA, Cho SJ et al. Mendelian randomization analysis of the effect of maternal homocysteine during pregnancy, as represented by maternal MTHFR C677T genotype, on birth weight. *J Epidemiol* 2013;23:371-375.
(75) Yang Q, Bailey L, Clarke R et al. Prospective study of methylenetetrahydrofolate reductase (MTHFR) variant C677T and risk of all-cause and cardiovascular disease mortality among 6000 US adults. *Am J Clin Nutr* 2012;95:1245-1253.
(76) Clarke R, Bennett DA, Parish S et al. Homocysteine and coronary heart disease: meta-analysis of MTHFR case-control studies, avoiding publication bias. *PLoS Med* 2012;9:e1001177.
(77) Marjot T, Yadav S, Hasan N, Bentley P, Sharma P. Genes associated with adult cerebral venous thrombosis. *Stroke* 2011;42:913-918.
(78) Bentley P, Peck G, Smeeth L, Whittaker J, Sharma P. Causal relationship of susceptibility genes to ischemic stroke: comparison to ischemic heart disease and biochemical determinants. *PLoS One* 2010;5:e9136.
(79) Nagele P, Zeugswetter B, Wiener C et al. Influence of methylenetetrahydrofolate reductase gene polymorphisms on homocysteine concentrations after nitrous oxide anesthesia. *Anesthesiology* 2008;109:36-43.
(80) Casas JP, Bautista LE, Smeeth L, Sharma P, Hingorani AD. Homocysteine and stroke: evidence on a causal link from mendelian randomisation. *Lancet* 2005;365:224-232.

(81) Huang T, Ren J, Huang J, Li D. Association of homocysteine with type 2 diabetes: a meta-analysis implementing Mendelian randomization approach. *BMC Genomics* 2013;14:867.
(82) Mamasoula C, Prentice RR, Pierscionek T et al. Association between C677T polymorphism of methylene tetrahydrofolate reductase and congenital heart disease: meta-analysis of 7697 cases and 13,125 controls. *Circ Cardiovasc Genet* 2013;6:347-353.
(83) Binder AM, Michels KB. The causal effect of red blood cell folate on genome-wide methylation in cord blood: a Mendelian randomization approach. *BMC Bioinformatics* 2013;14:353.
(84) Thuesen BH, Husemoen LLN, Fenger M, Linneberg A. Lack of association between the MTHFR (C677T) polymorphism and atopic disease. *Clinical Respiratory Journal* 2009;3:102-108.
(85) Granell R, Heron J, Lewis S, Davey Smith G, Sterne JAC, Henderson J. The association between mother and child MTHFR C677T polymorphisms, dietary folate intake and childhood atopy in a population-based, longitudinal birth cohort. *Clinical and Experimental Allergy* 2008;38:320-328.
(86) Ference BA, Yoo W, Alesh I et al. Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: a Mendelian randomization analysis. *J Am Coll Cardiol* 2012;60:2631-2639.
(87) Varbo A, Benn M, Tybjærg-Hansen A, Jørgensen AB, Frikke-Schmidt R, Nordestgaard BG. Remnant cholesterol as a causal risk factor for ischemic heart disease. *J Am Coll Cardiol* 2013;61:427-436.
(88) Stender S, Frikke-Schmidt R, Benn M, Nordestgaard BG, Tybjærg-Hansen A. Low-density lipoprotein cholesterol and risk of gallstone disease: a Mendelian randomization study and meta-analyses. *J Hepatol* 2013;58:126-133.
(89) Benn M, Tybjærg-Hansen A, Stender S, Frikke-Schmidt R, Nordestgaard BG. Low-density lipoprotein cholesterol and the risk of cancer: a mendelian randomization study. *J Natl Cancer Inst* 2011;103:508-519.
(90) Linsel-Nitschke P, Götz A, Erdmann J et al. Lifelong reduction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of coronary artery disease--a Mendelian Randomisation study. *PLoS One* 2008;3:e2986.
(91) Niu W, Zhang X, Qi Y. Association of an apolipoprotein E polymorphism with circulating cholesterols and hypertension: a meta-based Mendelian randomization analysis. *Hypertens Res* 2012;35:434-440.
(92) Shah S, Casas JP, Drenos F et al. Causal relevance of blood lipid fractions in the development of carotid atherosclerosis: Mendelian randomization analysis. *Circ Cardiovasc Genet* 2013;6:63-72.
(93) Qin XY, Tian J, Fang K et al. [Mendelian randomization study of the relationship between high-density lipoprotein cholesterol and age-related macular degeneration]. *Beijing Da Xue Xue Bao* 2012;44:407-411.
(94) Voight BF, Peloso GM, Orho-Melander M et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet* 2012;380:572-580.
(95) Haase CL, Tybjærg-Hansen A, Qayyum AA, Schou J, Nordestgaard BG, Frikke-Schmidt R. LCAT, HDL cholesterol and ischemic cardiovascular disease: a Mendelian randomization study of HDL cholesterol in 54,500 individuals. *J Clin Endocrinol Metab* 2012;97:E248-E256.

(96) Chmielewski M, Verduijn M, Drechsler C et al. Low cholesterol in dialysis patients--causal factor for mortality or an effect of confounding? *Nephrol Dial Transplant* 2011;26:3325-3331.
(97) Trompet S, Jukema JW, Katan MB et al. Apolipoprotein e genotype, plasma cholesterol, and cancer: a Mendelian randomization study. *Am J Epidemiol* 2009;170:1415-1421.
(98) Giltay EJ, van Reedt Dortland AK, Nissinen A et al. Serum cholesterol, apolipoprotein E genotype and depressive symptoms in elderly European men: the FINE study. *J Affect Disord* 2009;115:471-477.
(99) Jørgensen AB, Frikke-Schmidt R, West AS, Grande P, Nordestgaard BG, Tybjærg-Hansen A. Genetically elevated non-fasting triglycerides and calculated remnant cholesterol as causal risk factors for myocardial infarction. *Eur Heart J* 2013;34:1826-1833.
(100) Yao WM, Zhang HF, Zhu ZY et al. Genetically elevated levels of circulating triglycerides and brachial-ankle pulse wave velocity in a Chinese population. *J Hum Hypertens* 2013;27:265-270.
(101) Sarwar N, Sandhu MS, Ricketts SL et al. Triglyceride-mediated pathways and coronary disease: collaborative analysis of 101 studies. *Lancet* 2010;375:1634-1639.
(102) De Silva NM, Freathy RM, Palmer TM et al. Mendelian randomization studies do not support a role for raised circulating triglyceride levels influencing type 2 diabetes, glucose levels, or insulin resistance. *Diabetes* 2011;60:1008-1018.
(103) Thanassoulis G, Campbell CY, Owens DS et al. Genetic associations with valvular calcification and aortic stenosis. *N Engl J Med* 2013;368:503-512.
(104) Kamstrup PR, Tybjærg-Hansen A, Nordestgaard BG. Genetic evidence that lipoprotein(a) associates with atherosclerotic stenosis rather than venous thrombosis. *Arterioscler Thromb Vasc Biol* 2012;32:1732-1741.
(105) Kivimäki M, Magnussen CG, Juonala M et al. Conventional and Mendelian randomization analyses suggest no association between lipoprotein(a) and early atherosclerosis: the Young Finns Study. *Int J Epidemiol* 2011;40:470-478.
(106) Kamstrup PR, Tybjærg-Hansen A, Steffensen R, Nordestgaard BG. Genetically elevated lipoprotein(a) and increased risk of myocardial infarction. *JAMA* 2009;301:2331-2339.
(107) Kamstrup PR, Nordestgaard BG. Lipoprotein(a) concentrations, isoform size, and risk of type 2 diabetes: a Mendelian randomisation study. *The Lancet Diabetes & Endocrinology* 2013;1:220-227.
(108) Kamstrup PR, Tybjærg-Hansen A, Nordestgaard BG. Elevated lipoprotein(a) and risk of aortic valve stenosis in the general population. *J Am Coll Cardiol* 2014;63:470-477.
(109) Ye Z, Haycock PC, Gurdasani D et al. The association between circulating lipoprotein(a) and type 2 diabetes: is it causal? *Diabetes* 2014;63:332-342.
(110) Casas JP, Ninio E, Panayiotou A et al. PLA2G7 Genotype, lipoprotein-associated phospholipase A2 activity, and coronary heart disease risk in 10 494 cases and 15 624 controls of european ancestry. *Circulation* 2010;121:2284-2293.
(111) Koshy B, Miyashita A, St.Jean P et al. Genetic deficiency of plasma lipoprotein-associated phospholipase A 2 (PLA2G7 V297F null mutation) and risk of Alzheimer's disease in Japan. *Journal of Alzheimer's Disease* 2010;21:775-780.
(112) Holmes MV, Simon T, Exeter HJ et al. Secretory phospholipase A(2)-IIA and cardiovascular disease: a mendelian randomization study. *J Am Coll Cardiol* 2013;62:1966-1976.

(113) Dai X, Yuan J, Yao P et al. Association between serum uric acid and the metabolic syndrome among a middle- and old-age Chinese population. *Eur J Epidemiol* 2013;28:669-676.
(114) Parsa A, Brown E, Weir MR et al. Genotype-based changes in serum uric acid affect blood pressure. *Kidney Int* 2012;81:502-507.
(115) Hughes K, Flynn T, de Zoysa J, Dalbeth N, Merriman TR. Mendelian randomization analysis associates increased serum urate, due to genetic variation in uric acid transporters, with improved renal function. *Kidney Int* 2014;85:344-351.
(116) Pfister R, Barnes D, Luben R et al. No evidence for a causal link between uric acid and type 2 diabetes: a Mendelian randomisation approach. *Diabetologia* 2011;54:2561-2569.
(117) Niu W, Liu Y, Qi Y, Wu Z, Zhu D, Jin W. Association of interleukin-6 circulating levels with coronary artery disease: a meta-analysis implementing mendelian randomization approach. *Int J Cardiol* 2012;157:243-252.
(118) van Durme YM, Lahousse L, Verhamme KM et al. Mendelian randomization study of interleukin-6 in chronic obstructive pulmonary disease. *Respiration* 2011;82:530-538.
(119) Harrison SC, Smith AJ, Jones GT et al. Interleukin-6 receptor pathways in abdominal aortic aneurysm. *Eur Heart J* 2013;34:3707-3716.
(120) Hingorani AD, Casas JP. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. *Lancet* 2012;379:1214-1224.
(121) Jensen MK, Bartz TM, Djousse L et al. Genetically elevated fetuin-A levels, fasting glucose levels, and risk of type 2 diabetes: the cardiovascular health study. *Diabetes Care* 2013;36:3121-3127.
(122) Verduijn M, Prein RA, Stenvinkel P et al. Is fetuin-A a mortality risk factor in dialysis patients or a mere risk marker? A Mendelian randomization approach. *Nephrol Dial Transplant* 2011;26:239-245.
(123) Fisher E, Stefan N, Saar K et al. Association of AHSG gene polymorphisms with fetuin-A plasma levels and cardiovascular diseases in the EPIC-Potsdam study. *Circ Cardiovasc Genet* 2009;2:607-613.
(124) Yaghootkar H, Lamina C, Scott RA et al. Mendelian randomization studies do not support a causal role for reduced circulating adiponectin levels in insulin resistance and type 2 diabetes. *Diabetes* 2013;62:3589-3598.
(125) Mente A, Meyre D, Lanktree MB et al. Causal relationship between adiponectin and metabolic traits: a Mendelian randomization study in a multiethnic population. *PLoS One* 2013;8:e66808.
(126) Gao H, Fall T, van Dam RM et al. Evidence of a causal relationship between adiponectin levels and insulin sensitivity: a Mendelian randomization study. *Diabetes* 2013;62:1338-1344.
(127) Menzaghi C, De Cosmo S, Copetti M et al. Relationship between ADIPOQ gene, circulating high molecular weight adiponectin and albuminuria in individuals with normal kidney function: Evidence from a family-based study. *Diabetologia* 2011;54:812-818.
(128) Klovaite J, Nordestgaard BG, Tybjærg-Hansen A, Benn M. Elevated fibrinogen levels are associated with risk of pulmonary embolism, but not with deep venous thrombosis. *Am J Respir Crit Care Med* 2013;187:286-293.

(129) Keavney B, Danesh J, Parish S et al. Fibrinogen and coronary heart disease: test of causality by 'Mendelian randomization'. *Int J Epidemiol* 2006;35:935-943.
(130) Davey Smith G, Harbord R, Milton J, Ebrahim S, Sterne JAC. Does elevated plasma fibrinogen increase the risk of coronary heart disease? Evidence from a meta-analysis of genetic association studies. *Arteriosclerosis, Thrombosis, and Vascular Biology* 2005;25:2228-2233.
(131) Bonilla C, Lawlor DA, Ben-Shlomo Y et al. Maternal and offspring fasting glucose and type 2 diabetes-associated genetic variants and cognitive function at age 8: a Mendelian randomization study in the Avon Longitudinal Study of Parents and Children. *BMC Med Genet* 2012;13:90.
(132) Rasmussen-Torvik LJ, Li M, Kao WH et al. Association of a fasting glucose genetic risk score with subclinical atherosclerosis: The Atherosclerosis Risk in Communities (ARIC) study. *Diabetes* 2011;60:331-335.
(133) Trombetta M, Bonetti S, Boselli ML et al. PPARG2 Pro12Ala and ADAMTS9 rs4607103 as "insulin resistance loci" and "insulin secretion loci" in Italian individuals. The GENFIEV study and the Verona Newly Diagnosed Type 2 Diabetes Study (VNDS) 4. *Acta Diabetol* 2013;50:401-408.
(134) Song Y, Yeung E, Liu A et al. Pancreatic beta-cell function and type 2 diabetes risk: quantify the causal effect using a Mendelian randomization approach based on meta-analyses. *Hum Mol Genet* 2012;21:5010-5018.
(135) Benn M, Tybjærg-Hansen A, McCarthy MI, Jensen GB, Grande P, Nordestgaard BG. Nonfasting glucose, ischemic heart disease, and myocardial infarction: a Mendelian randomization study. *J Am Coll Cardiol* 2012;59:2356-2365.
(136) Ioannidis A, Ikonomi E, Dimou NL, Douma L, Bagos PG. Polymorphisms of the insulin receptor and the insulin receptor substrates genes in polycystic ovary syndrome: a Mendelian randomization meta-analysis. *Mol Genet Metab* 2010;99:174-183.
(137) Pierce BL, Ahsan H. Genetic susceptibility to type 2 diabetes is associated with reduced prostate cancer risk. *Human Heredity* 2010;69:193-201.
(138) Almon R, Álvarez-León EE, Serra-Majem L. Association of the European lactase persistence variant (LCT-13910 C>T polymorphism) with obesity in the Canary Islands. *PLoS One* 2012;7:e43978.
(139) Almon R, Álvarez-León EE, Engfeldt P, Serra-Majem L, Magnuson A, Nilsson TK. Associations between lactase persistence and the metabolic syndrome in a cross-sectional study in the Canary Islands. *Eur J Nutr* 2010;49:141-146.
(140) Travis RC, Appleby PN, Siddiq A et al. Genetic variation in the lactase gene, dairy product intake and risk for prostate cancer in the European prospective investigation into cancer and nutrition. *International Journal of Cancer* 2013;132:1901-1910.
(141) Pichler I, Del Greco MF, Gögele M et al. Serum iron levels and the risk of Parkinson disease: a Mendelian randomization study. *PLoS Med* 2013;10:e1001462.
(142) Alwan NA, Lawlor DA, McArdle HJ, Greenwood DC, Cade JE. Exploring the relationship between maternal iron status and offspring's blood pressure and adiposity: a Mendelian randomization study. *Clin Epidemiol* 2012;4:193-200.
(143) Gan W, Guan Y, Wu Q et al. Association of TMPRSS6 polymorphisms with ferritin, hemoglobin, and type 2 diabetes risk in a Chinese Han population. *Am J Clin Nutr* 2012;95:626-632.

(144) Stender S, Frikke-Schmidt R, Nordestgaard BG, Tybjærg-Hansen A. Extreme bilirubin levels as a causal risk factor for symptomatic gallstone disease. *JAMA Intern Med* 2013;173:1222-1228.
(145) Stender S, Frikke-Schmidt R, Nordestgaard BG, Grande P, Tybjærg-Hansen A. Genetically elevated bilirubin and risk of ischaemic heart disease: three Mendelian randomization studies and a meta-analysis. *J Intern Med* 2013;273:59-68.
(146) McArdle PF, Whitcomb BW, Tanner K, Mitchell BD, Shuldiner AR, Parsa A. Association between bilirubin and cardiovascular disease risk factors: using Mendelian randomization to assess causal inference. *BMC Cardiovasc Disord* 2012;12:16.
(147) Perry JR, Weedon MN, Langenberg C et al. Genetic evidence that raised sex hormone binding globulin (SHBG) levels reduce the risk of type 2 diabetes. *Hum Mol Genet* 2010;19:535-544.
(148) Ding EL, Song Y, Manson JE et al. Sex hormone-binding globulin and risk of type 2 diabetes in women and men. *N Engl J Med* 2009;361:1152-1163.
(149) Stegeman BH, Helmerhorst FM, Vos HL, Rosendaal FR, Van Hylckama Vlieg A. Sex hormone-binding globulin levels are not causally related to venous thrombosis risk in women not using hormonal contraceptives. *Journal of Thrombosis and Haemostasis* 2012;10:2061-2067.
(150) Haring R, Teumer A, Völker U et al. Mendelian randomization suggests non-causal associations of testosterone with cardiometabolic risk factors and mortality. *Andrology* 2013;1:17-23.
(151) Zhao J, Jiang C, Lam TH et al. Genetically predicted testosterone and cardiovascular risk factors in men: a Mendelian randomization analysis in the Guangzhou Biobank Cohort Study. *Int J Epidemiol* 2014;43:140-148.
(152) Attermann J, Obel C, Bilenberg N, Nordenbæk CM, Skytthe A, Olsen J. Traits of ADHD and autism in girls with a twin brother: a Mendelian randomization study. *Eur Child Adolesc Psychiatry* 2012;21:503-509.
(153) Guessous I, Dobrinas M, Kutalik Z et al. Caffeine intake and CYP1A2 variants associated with high caffeine intake protect non-smokers from hypertension. *Hum Mol Genet* 2012;21:3283-3292.
(154) Bech BH, Autrup H, Nohr EA, Henriksen TB, Olsen J. Stillbirth and slow metabolizers of caffeine: comparison by genotypes. *Int J Epidemiol* 2006;35:948-953.
(155) Bonilla C, Lawlor DA, Taylor AE et al. Vitamin B-12 status during pregnancy and child's IQ at age 8: a Mendelian randomization study in the Avon longitudinal study of parents and children. *PLoS One* 2012;7:e51084.
(156) Collin SM, Metcalfe C, Palmer TM et al. The causal roles of vitamin B(12) and transcobalamin in prostate cancer: can Mendelian randomization analysis provide definitive answers? *Int J Mol Epidemiol Genet* 2011;2:316-327.
(157) Bjørngaard JH, Gunnell D, Elvestad MB et al. The causal role of smoking in anxiety and depression: a Mendelian randomization analysis of the HUNT study. *Psychol Med* 2013;43:711-719.
(158) Lewis SJ, Araya R, Davey Smith G et al. Smoking is associated with, but does not cause, depressed mood in pregnancy--a mendelian randomization study. *PLoS One* 2011;6:e21689.

(159) Scott JA, Berkley JA, Mwangi I et al. Relation between falciparum malaria and bacteraemia in Kenyan children: a population-based, case-control study and a longitudinal study. *Lancet* 2011;378:1316-1323.
(160) Kang H, Kreuels B, Adjei O, Krumkamp R, May J, Small DS. The causal effect of malaria on stunting: a Mendelian randomization and matching approach. *International Journal of Epidemiology* 2013;42:1390-1398.
(161) Frayling TM, Rafiq S, Murray A et al. An interleukin-18 polymorphism is associated with reduced serum concentrations and better physical functioning in older people. *J Gerontol A Biol Sci Med Sci* 2007;62:73-78.
(162) Herder C, Klopp N, Baumert J et al. Effect of macrophage migration inhibitory factor (MIF) gene variants and MIF serum concentrations on the risk of type 2 diabetes: results from the MONICA/KORA Augsburg Case-Cohort Study, 1984-2002. *Diabetologia* 2008;51:276-284.
(163) Bouthoorn SH, van Lenthe FJ, Kiefte-de Jong JC et al. Genetic taste blindness to bitter and body composition in childhood: a Mendelian randomization design. *Int J Obes (Lond)* 2014;38:1005-1010.
(164) You NC, Chen BH, Song Y et al. A prospective study of leukocyte telomere length and risk of type 2 diabetes in postmenopausal women. *Diabetes* 2012;61:2998-3004.
(165) Breitling LP, Koenig W, Fischer M et al. Type II secretory phospholipase A2 and prognosis in patients with stable coronary heart disease: mendelian randomization study. *PLoS One* 2011;6:e22318.
(166) Conen D, Vollenweider P, Rousson V et al. Use of a Mendelian randomization approach to assess the causal relation of gamma-Glutamyltransferase with blood pressure and serum insulin levels. *Am J Epidemiol* 2010;172:1431-1441.
(167) Kröger J, Zietemann V, Enzenbach C et al. Erythrocyte membrane phospholipid fatty acids, desaturase activity, and dietary fatty acids in relation to risk of type 2 diabetes in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam Study. *Am J Clin Nutr* 2011;93:127-142.
(168) Love-Gregory L, Sherva R, Schappe T et al. Common CD36 SNPs reduce protein expression and may contribute to a protective atherogenic profile. *Hum Mol Genet* 2011;20:193-201.
(169) Wu Y, Li H, Loos RJ et al. RBP4 variants are significantly associated with plasma RBP4 levels and hypertriglyceridemia risk in Chinese Hans. *J Lipid Res* 2009;50:1479-1486.
(170) Sharma NK, Gupta A, Prabhakar S et al. Association between CFH Y402H polymorphism and age related macular degeneration in North Indian cohort. *PLoS One* 2013;8:e70193.
(171) Aramini B, Kim C, Diangelo S et al. Donor surfactant protein D (SP-D) polymorphisms are associated with lung transplant outcome. *American Journal of Transplantation* 2013;13:2130-2136.
(172) Tian Q, Jia J, Ling S, Liu Y, Yang S, Shao Z. A causal role for circulating miR-34b in osteosarcoma. *Eur J Surg Oncol* 2014;40:67-72.
(173) Oelsner EC, Pottinger TD, Burkart KM et al. Adhesion molecules, endothelin-1 and lung function in seven population-based cohorts. *Biomarkers* 2013;18:196-203.
(174) Cruchaga C, Kauwe JS, Nowotny P et al. Cerebrospinal fluid APOE levels: an endophenotype for genetic studies for Alzheimer's disease. *Hum Mol Genet* 2012;21:4558-4571.

(175) Pfister R, Sharp S, Luben R et al. Mendelian randomization study of B-type natriuretic peptide and type 2 diabetes: evidence of causal association from population studies. *PLoS Med* 2011;8:e1001112.
(176) Perry JR, Ferrucci L, Bandinelli S et al. Circulating beta-carotene levels and type 2 diabetes-cause or effect? *Diabetologia* 2009;52:2117-2121.
(177) Pierce BL, Tong L, Argos M et al. Arsenic metabolism efficiency has a causal role in arsenic toxicity: Mendelian randomization and gene-environment interaction. *International Journal of Epidemiology* 2013;42:1862-1871.
(178) Adamsson Eryd S, Sjögren M, Smith JG et al. Ceruloplasmin and atrial fibrillation: evidence of causality from a population-based Mendelian randomization study. *J Intern Med* 2014;275:164-171.
(179) Rice NE, Bandinelli S, Corsi AM et al. The paraoxonase (PON1) Q192R polymorphism is not associated with poor health status or depression in the ELSA or INCHIANTI studies. *Int J Epidemiol* 2009;38:1374-1379.

Chapter 8

Mendelian randomization studies in the elderly.

Anna G.C. Boef, Saskia le Cessie and Olaf M. Dekkers

Epidemiology. 2015 Mar;26(2):e15-6

Mendelian randomization studies use genetically determined variation in exposure levels to study causal effects. Unmeasured confounding and reverse causation, which hamper the estimation of causal effects of exposures in conventional analysis, can thereby be circumvented.[1] The genetic variant used as an instrument should fulfil the following conditions: (1) it is associated with the exposure; (2) it only affects the outcome through the exposure; and (3) it is not related to other factors which affect the outcome.[1-3] If these main assumptions are fulfilled, the effect of the genetic variant on the outcome can be attributed to the effect of the exposure on the outcome. An overview of numerous scenarios in which these assumptions are violated was published recently, with a discussion of the consequences and suggestions for alternative approaches to Mendelian randomization studies.[3]

Another scenario in which the assumptions could be violated is if a Mendelian randomization study is performed in an elderly population. Let us think of a Mendelian randomization study in which the effect of exposure X on outcome Y is studied in a population aged > 80 years, using genetic variant G as the genetic instrument to circumvent the confounding by unmeasured factors U_1 of the X-Y relation. If the genetic variant G, through its effect on the exposure X has affected survival up to age 80, collider stratification bias (selection bias) may occur. This is illustrated in the Figure: both genotype G (through exposure X) and other risk factors U_2 affect survival up to age 80. Survival S is therefore a collider and restriction of the population to those who have survived up to age 80 results in collider stratification bias by inducing an association between G and U_2. The intuitive interpretation of this phenomenon is as follows. We assume that genotype G increases mortality rates. If a person with a genotype G is still alive at age 80, this person will be less likely to have other risk factors for mortality (high blood pressure, smoking, etc.) compared to people without genotype G. This means that in the population aged over 80 the genetic variant is associated with other factors which affect the outcome, violating assumption 2. The effect of the genetic variant G on the outcome Y can therefore no longer be solely attributed to the effect of the exposure X on outcome Y.

An example in which this collider stratification bias might occur is a Mendelian randomization study in subjects aged over 80 using *APOE* variants (G) as an instrument for cholesterol level (X), with (for example) myocardial infarction as the outcome (Y). *APOE* variants are known to cause variation in cholesterol levels. Because cholesterol levels will have influenced survival up to age 80 (S) and hence selection into the study population, an association of *APOE* variants with other risk factors which influenced survival up to age 80 (U_2, e.g. smoking) is introduced. For example, among those with a cholesterol increasing *APOE*-variant who have survived up to age 80 there will be fewer smokers than among those without the variant. If these other factors also affect the risk of myocardial infarction, the genotype-outcome relation will be biased: in this

example the bias in the estimated effect of genetically increased cholesterol will be towards a lower risk of myocardial infarction due to the inverse relation with smoking. This also applies if the outcome is survival (from age 80).
The bias introduced by selection on survival will of course be most prominent for Mendelian randomization studies investigating exposures which strongly affect survival.

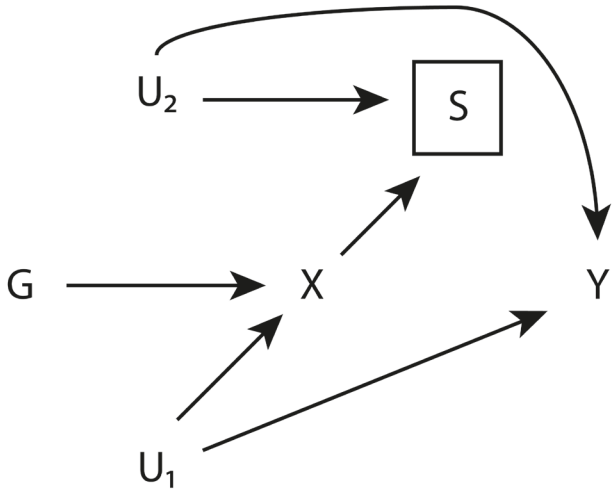


Figure. *Genetic variant G is an instrument for the effect of exposure X on outcome Y, with unmeasured factors U_1 confounding the X-Y relation. Both X and unmeasured risk factors U_2 affect survival S, and U_2 also affects Y. Because of selection on S, there is an association between G and U_2.*

Reference List

(1) Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* 2008;27:1133-1163.
(2) Swanson SA, Hernan MA. Commentary: how to report instrumental variable analyses (suggestions welcome). *Epidemiology* 2013;24:370-374.
(3) VanderWeele TJ, Tchetgen Tchetgen EJ, Cornelis M, Kraft P. Methodological challenges in mendelian randomization. *Epidemiology* 2014;25:427-435.

Chapter 9

General discussion

In this thesis we aimed to investigate the validity of instrumental variable analysis, in particular using physician's preference as an instrument, to evaluate beneficial or adverse effects of interventions and we aimed to identify the settings and types of questions for which it most useful. In this chapter we present a summary of our main findings, discuss strengths and limitations of our research and consider its implications.

Summary of the principal findings

In **Chapter 2** we found substantial variation amongst general practitioners in their treatment decisions when presented with the same set of eight fictitious cases of patients with subclinical hypothyroidism, supporting the existence of physician's prescribing preference. Further, we found that the *deterministic* monotonicity assumption (that the instrument is related to treatment monotonically in one direction for all patients) did not hold even in a relatively simple setting, suggesting that this assumption is not plausible when physician's preference is used as an instrumental variable. In contrast, we found that the *stochastic* monotonicity assumption (that the instrument is related to treatment monotonically across subjects within strata of a sufficient set of measured and unmeasured common causes of treatment and the outcome) held in the survey data when a different 'prescription' of the same general practitioner was used as the instrument. This suggests that a more relaxed version of the monotonicity assumption may be plausible when physician's preference is used as an instrumental variable.

We subsequently applied physician's preference based instrumental variable analysis in a clinical epidemiological study of typical size (a few hundred patients) in **Chapter 3**. We found that estimates of the effect of preoperative corticosteroids on mechanical ventilation time, duration of intensive care and hospital stay, occurrence of infections, atrial fibrillation, heart failure and delirium in elective cardiac surgery patients were similar in direction to estimates from a randomised controlled trial. However, the estimated effects were much larger with uninformative wide confidence intervals. We concluded that the lesser statistical precision of instrumental variable analysis limits its usefulness in a study that might be of sufficient size for conventional analyses - even if a strong and plausible instrument is available.

In **Chapter 4** we showed through simulations how the performance of instrumental variable analysis in comparison to conventional analyses, a bias-variance trade-off, depends substantially on sample size. Other determinants are the strength of the instrument and the strength of confounding. We derived an equation that can be used to approximate a 'threshold' sample size above which the mean squared error (a summary measure of bias and variance) of instrumental variable analyses will be lower than that of conventional analyses. Further, we showed that substantial sample

sizes will generally be needed for the bias-variance trade-off to be in favour of an instrumental variable analysis in epidemiologic studies.

In **Chapter 5**, we investigated whether instrumental variable analysis is useful as a sensitivity analysis in studies of adverse effects to assess the presence of confounding. The topic of the study was the comparison of the occurrence of venous thrombosis in users of third generation oral contraceptives vs. second generation contraceptives. In principle, this is an unpredictable unintended effect and we therefore investigated whether an instrumental variable analysis would yield the same estimates as an analysis using standard statistical methods to adjust for confounding (as in principle we would expect little confounding). The study population consisted of new users of second or third generation oral contraceptives, derived from a very large primary care database. We showed that the instrumental variable estimates (using general practitioner's preference as an instrument) of the effect of third versus second generation oral contraceptives on occurrence of venous thromboembolism were similar to estimates from conventional analyses. If anything, the conventional analysis seemed to be more conservative than the instrumental variable analysis. However, even in this very large study population the variance of the instrumental variable estimates was very large, due to the relatively rare outcome. Further, the analysis was complicated because changes in both prescribing preference and patient characteristics over time resulted in violation of the independence assumption and necessitated an adjusted instrumental variable analysis. We concluded that major confounding was unlikely due to the similarity of the estimates obtained under different sets of assumptions.

In **Chapter 6** we recommended to report the association between the instrument and the outcome before performing any formal instrumental variable analysis, which amounts to a conventional epidemiologic analysis. This is important because it shows the association which is subsequently extrapolated in the formal instrumental variable analysis. This recommendation was proposed in response to a paper outlining guidelines for the reporting of instrumental variable analysis.

We then shifted our focus from physician's preference based instrumental variable analysis to Mendelian randomisation studies. In **Chapter 7** we reviewed methodological approaches used in Mendelian randomisation studies and found that the specific methods used vary widely, falling broadly into three categories: 1) using genetic information as a proxy for the exposure without further estimation, 2) performing an instrumental variable analysis; 3) comparing the observed with the expected genotype-outcome association. Further we found that Mendelian randomisation studies often insufficiently discuss underlying assumptions and

report statistical methods for IV analysis. In a certain sense the fact that Mendelian randomisation studies are also instrumental variable studies (with or without a formal instrumental variable analysis) is often disregarded. We therefore devised a checklist for the reporting of Mendelian randomisation studies. Finally, in **Chapter 8** we explained that collider-stratification bias may exist if Mendelian randomisation studies are performed in elderly populations, as both the genetic variant used as an instrument and other causes of the outcome may be causally related to survival up to the age at which the population is selected.

Implications and recommendations

Instrumental variable analysis as a primary analysis

Instrumental variable analysis for estimation of therapeutic effects has the greatest potential in situations in which very large datasets are available and in which there is substantial confounding by indication and also little available information on these confounding factors (see conclusion of Chapter 4). However, if the direction of this confounding is predictable, sensitivity analyses could alternatively be used to derive a plausible range of the treatment effect.[1-3] Instrumental variable analysis would therefore more specifically be suited to situations in which there is confounding with such complexity that the direction and magnitude of the resulting bias is unpredictable. Further we remark that although instrumental variable analysis is of most value in situations with substantial unmeasured confounding, paradoxically, substantial unmeasured confounding limits the potential strength of any instrument (as this confounding will determine a substantial part of the variation in exposure).[4]

Instrumental variable analysis as a sensitivity analysis

In case of unpredictable adverse effects, such as the increased risk of venous thromboembolism of 3rd generation oral contraceptives in comparison to 2nd generation oral contraceptives, confounding by (contra-)indication[5] is unlikely and instrumental variable analysis would therefore not be particularly suitable as a primary analysis. However, it may be used as a sensitivity analysis: if results of an instrumental variable analysis are similar to those of the conventional analyses, this supports the notion that there is little confounding by contra-indication (provided, of course, that instrumental variable assumptions hold, and that suitably large databases exist). Instrumental variable analysis may also have a role as a sensitivity analysis in studies of intended effects.[6] Looking at the same data under different sets of assumptions can contribute to the understanding of causal effects.[7]

Types of therapeutic question for which instrumental variable analysis is most useful

Instrumental variable analysis for estimation of therapeutic effects seems primarily

suited to therapeutic decisions in which there are two clearly defined alternatives. Davies et al compare 5 different treatment options (2 selective COX-2 inhibitors and 3 non-specific NSAIDs) as a sensitivity analysis in a study primarily comparing the 2 drug classes.[8] However, examples of instrumental variable studies in which more than two treatment options are compared are rare, although there are many situations in which there are more than two alternatives to consider in a treatment decision. Even if only two options are primarily of interest for the research question at hand, exclusion of patients who received other treatment options may result in inclusion of different subsets of the patient population for different physicians, threatening the validity of the independence assumption (i.e. that there is no confounding of the instrument-outcome relation). This may have been an issue in **Chapter 5**, in which the introduction of drospirenone-containing oral contraceptives resulted in an additional option besides 2nd and 3rd generation oral contraceptives. Whereas formerly the comparison of 2nd and 3rd generation oral contraceptives will have included (nearly) all women who started using a combined hormonal oral contraceptive, the introduction of an additional option will have reduced this population in a manner which is not necessarily random.

Instrumental variable analysis of randomised controlled trials

In the ideal randomised controlled trial with complete compliance and complete follow-up, the treatment effect estimated is the average treatment effect in the population. This changes when compliance is incomplete, in which case an intention-to-treat effect is usually estimated. This estimates the effect of assigning the treatment rather than the effect of taking the treatment, and is therefore not always the effect of interest.[9;10] The intention-to-treat effect is a conservative estimate of the effect of taking treatment, i.e. biased towards the null. This can be particularly problematic in studies of adverse effects or in non-inferiority trials.[11] An as-treated analysis or a per-protocol analysis on the other hand essentially negates the randomisation, resulting in incomparable populations and an estimate without a clear causal interpretation.[9;11] The analysis of RCTs with non-compliance using methods common to observational studies has been advocated, in order to obtain a valid estimate of the effect of taking treatment (besides the intention-to-treat estimate of the effect of assigned treatment).[11] One such way is to perform an instrumental variable analysis, using treatment arm as an instrumental variable. This will usually be a strong instrumental variable, as treatment arm strongly predicts treatment unless compliance is very low.[12] In the context of a randomised trial, the deterministic monotonicity assumption, i.e. the absence of defiers (subjects who would take the opposite of what they are assigned to in either treatment arm) is usually reasonable. Under this assumption the instrumental variable analysis estimates the effect within the compliers: those who take the treatment to which they

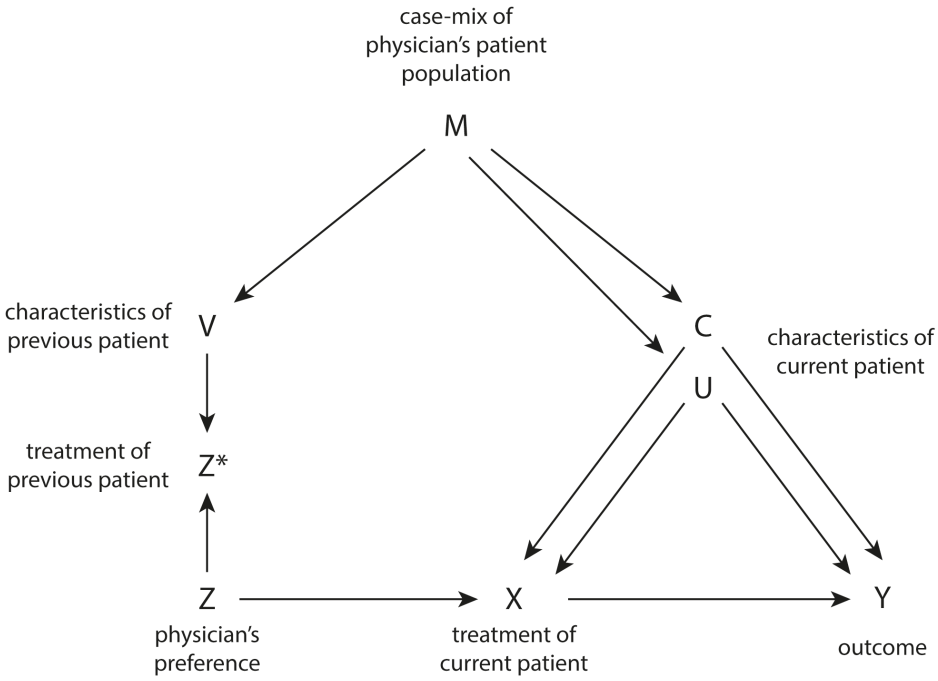
are assigned. Importantly, the interpretation of this estimate therefore differs from the average treatment effect in the population and from the intention-to-treat effect. The estimate can for example be obtained using the Wald estimator, which divides the intention-to-treat effect by the difference in probability of being treated with the study treatment.[10;12] For time-varying treatments (time-varying adherence) more complex instrumental variable methods such as g-estimation can be used.[11;13] Examples of instrumental variable analysis in RCTs are as yet rare. One example is a randomised trial investigating the effect of yoga (in addition to usual general practitioner care) on chronic low back pain. The intention-to-treat estimate of the effect of assignment to yoga in addition to usual care on 3-month Roland Morris Disability Questionnaire score was -2.17 (95% CI -3.31; -1.03), whereas the complier-average-causal-effect estimate of attending at least one yoga class was -2.45 (-3.67;-1.24) and the complier-average-causal-effect estimate of attending all 12 offered yoga classes was -3.30 (-4.90;-1.70).[14]

Physician's preference as an instrumental variable

Physician's preference can specifically be useful as an instrumental variable in situations in which there are no stringently applied medical practice guidelines, i.e. when there is room for preference to play a role in treatment decisions. An example of such a situation is the decision whether to treat patients with subclinical hypothyroidism. The Dutch general practitioners' guideline, for example, does not recommend treatment of subclinical hypothyroidism in general, but states that general practitioners may choose to try levothyroxine treatment and evaluate whether symptoms improve.[15] We showed in **Chapter 2** that there was substantial variation in treatment decisions among general practitioners presented with the same subclinical hypothyroidism cases. In these survey data, this variation remained after adjusting for characteristics of the GP and their patient population, which was reassuring with regard to the main instrumental variable assumptions.

However, a note of caution is warranted with regard to the exclusion restriction assumption. In applications of physician's preference based instrumental variable analysis, the treatment choice by the physician for (one or more) previous patients is usually used as an estimate of physician's preference, because this preference is not a directly measurable characteristic. Yet situations may occur in which instrumental variable assumptions hold for the underlying preference, but not for *an estimate* of this preference based on previous prescriptions.[16] We will explain this using the directed acyclic graph depicted in Figure 1. In this figure the treatment of the previous patient Z* is used as a proxy for underlying preference of Z. If there is variation in the case-mix M of the different physicians in the study, the current patient and the previous patient

are likely to be more similar than any random two patients from the entire study population. The instrument Z* will then be related to characteristics of the current patient C and U, through case-mix M and characteristics of the previous patient V, violating the independence assumption. Note that underlying preference Z is still a valid instrument, because Z* acts as a collider, blocking the path between Z and patient characteristics C and U. In the setting of **Chapter 3**, we think this is unlikely to have occurred, as patients will have been treated by the anaesthesiologist on duty and systematic differences in case-mix are therefore unlikely to exist. For studies in which preference of general practitioners is used as an instrument, differences in case-mix are likely (e.g. due to geographical variation in socio-economic status and demographic characteristics) and the confounding described may therefore threaten the validity of previous prescriptions of the GP as an instrument.



What does an instrumental variable analysis estimate?

As stated previously, in the ideal randomised controlled trial with complete compliance and complete follow-up, the treatment effect estimated is the average treatment effect in the study population. If the study population is representative of the population of interest, the estimate represents the average treatment effect in the population of interest. However, due to inclusion and exclusion criteria, RCT populations are often

not representative of the population of interest, which limits the generalisability of results. An often stated advantage of observational studies over randomised controlled trials (RCTs) is the greater generalisability of the results, because subjects who are unlikely to be included in RCTs can be part of the study population in an observational study. Conventional methods to adjust for confounding in observational studies results in estimates which represent average effects in the study population conditional on the confounders for which they have adjusted. For results of an instrumental variable study the question to whom the results apply and how to interpret the estimate is more complicated. It depends on the assumption which is made in order to obtain a point estimate (the "fourth" assumption, in addition to the three main instrumental variable assumptions). Under the assumption of homogeneity of treatment effects the results apply to the entire study population. However, if homogeneity of treatment effects is not realistic, another assumption is necessary in order to obtain an interpretable point estimate. This is often some form of the monotonicity assumption, in which case the question to which population the results apply is more complex.

The exact interpretation of the instrumental variable estimate depends on the version of the monotonicity assumption. The situation with a binary instrument and the *deterministic* monotonicity assumption is relatively easy: the point estimate represents the average effect among the 'compliers'. However, the compliers in the study population cannot be identified: all subjects only experienced the treatment they received at the actual value of the instrument and the treatment they would have received at the counterfactual value of the instrument is unknown. A description of the distribution of the characterisation of the compliers is possible however,[17;18] as described by Angrist and Pischke.[18]

Under the *stochastic* monotonicity assumption the question to which patients the results apply becomes more difficult, as the estimate is a weighted average of treatment effects. Clearly, the results do not apply to those subjects who evidently would have received the same treatment regardless of the value of the instrument (i.e. regardless of the preference of their physician – e.g. because of overriding medical reasons). The degree to which results apply to other patients depends not only to the proportion of the study population which consisted of that type of patient, but also on the strength of the instrument for the specific type of patient: the stronger the instrument, the higher the relative contribution of that specific type of patients to the estimate. A characterisation of the strength of instrumental variable weighted average treatment effect (SIVWATE) population would be rather more difficult than the characterisation of the compliers described previously.

In part because of the difficulty in the interpretation of the point estimate from an instrumental variable analysis and the additional assumption required, the reporting of bounds of the instrumental variable estimate has been advocated in reporting guidelines for instrumental variable analysis.[17] These bounds represent the upper and lower limits of the average causal effect in the study population.[19] Balke and Pearl describe the calculation of these bounds and how they can be narrowed under different assumptions.[20] In practice these bounds will generally be so disparate that they are uninformative. The main value in calculating bounds lies in the subsequent explicit decision on which fourth assumption is plausible and how the point estimate consequently should be interpreted.

Reflections on Mendelian randomisation and subsequent considerations for other forms of instrumental variable analysis

There is discussion on whether Mendelian randomisation should be viewed as instrumental variable analysis with genetic instrumental variables.[21;22] One point of discussion is that some studies qualify as Mendelian randomisation studies but do not perform a formal instrumental variable *analysis*.[22] More recently the question if and when formal instrumental variable analysis should be performed in Mendelian randomisation studies has been addressed by VanderWeele.[23] Most importantly this depends on the definition of the exposure and the consequences of this definition for the validity of the main instrumental variable assumptions: in some situations an estimate of the effect of the genetic instrument on the outcome may be valid, while no valid IV estimate of the effect of the exposure on the outcome can be obtained.[23] One example discussed by VanderWeele is that the effect of certain genetic variants on smoking behaviour will not be entirely captured by the effect on the number of cigarettes smoked per day: IV analysis estimating the effect of the number of cigarettes per day on lung cancer using these variants as instruments will then be biased because there are additional pathways from the variants to lung cancer (via other aspects of smoking behaviour). Investigating the association between the variants and lung cancer is a valid test of the presence of an effect of smoking behaviour in a more general sense on lung cancer.[23]

The question if and when a formal instrumental variable analysis should be performed applies not just to MR studies, but also to other instrumental variable studies. In case of studies of therapeutic effects, adequately capturing all aspects of the association between the instrument and the exposure may not be as difficult as in Mendelian randomisation studies. However, the outcome of interest in a study of therapeutic effects is often a binary or survival outcome. Formal IV analysis methods for such outcomes are not yet well-established and obtaining a valid quantitative estimate of

effect of the therapy on the outcome is therefore difficult, but estimation of the effect of the exposure on the outcome can serve as a test of causality. For many therapeutic questions quantification of the therapeutic effect will be important however, and without a formal instrumental variable analysis only the existence and direction of the effect can be evaluated.

Based on our findings in **Chapter 7** we would argue that for some aspects of Mendelian randomisation studies the instrumental variable perspective can be valuable. For example, reporting guidelines for instrumental variable analysis largely apply to Mendelian randomisation studies and can aid in improving the reporting of future MR studies. Furthermore, many Mendelian randomisation studies obtain an instrumental-variable type estimate of the effect of the exposure on the outcome using statistical methods which are not well-established (e.g. estimation of an odds ratio using a method similar to the Wald estimator). It is important that those who perform such studies are aware that the problems and limitations of instrumental variable analyses with binary or survival outcomes also apply to MR studies using these methods.

Conclusions

We set out to investigate the validity and usefulness of instrumental variable analysis, in particular using physician's preference as an instrument, in clinical epidemiological studies. We aimed to expose potential problems and limitations of this method and to identify the settings and types of questions for which it is most useful. By exploring several aspects of the method in both applications using existing data and in simulation studies we came to a number of conclusions. Instrumental variable analysis can be of value as a primary analysis in very large epidemiological studies with substantial unmeasured confounding (of unpredictable direction and magnitude) and a strong instrument. For studies of a more typical size in clinical epidemiology, e.g. several hundreds of patients an instrumental variable estimate will generally be uninformatively imprecise. A more broadly suitable role for instrumental variable analysis could be as a sensitivity analysis, for example to assess the presence of confounding (by contra-indication) in studies of adverse effects. Treatment preference differences exist between physicians, independent of characteristics of these physicians and their patient populations. Ascertaining that IV assumptions hold not only for underlying physician's preference, but also for the estimate of preference used as a proxy instrument, is paramount. For physician's preference as an instrument the stochastic monotonicity assumption may be a plausible assumption for obtaining a point estimate. Viewing Mendelian randomisation studies as instrumental variable studies can aid in improving the reporting of Mendelian randomisation studies, even if no formal instrumental variable analysis is performed.

Reference List

(1) Greenland S. Basic methods for sensitivity analysis of biases. *Int J Epidemiol* 1996;25:1107-1116.

(2) Arah OA, Chiba Y, Greenland S. Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. *Ann Epidemiol* 2008;18:637-646.

(3) Vanderweele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology* 2011;22:42-52.

(4) Martens EP, Pestman WR, de BA, Belitser SV, Klungel OH. Instrumental variables: application and limitations. *Epidemiology* 2006;17:260-267.

(5) Feenstra H, Grobbee RE, in't Veld BA, Stricker BH. Confounding by contraindication in a nationwide cohort study of risk for death in patients taking ibopamine. *Ann Intern Med* 2001;134:569-572.

(6) Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000;29:722-729.

(7) Glymour MM, Tchetgen Tchetgen EJ, Robins JM. Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *Am J Epidemiol* 2012;175:332-339.

(8) Davies NM, Smith GD, Windmeijer F, Martin RM. COX-2 selective nonsteroidal anti-inflammatory drugs and risk of gastrointestinal tract complications and myocardial infarction: an instrumental variable analysis. *Epidemiology* 2013;24:352-362.

(9) Sussman JB, Hayward RA. An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised controlled trials. *BMJ* 2010;340:c2073.

(10) Shrier I, Steele RJ, Verhagen E, Herbert R, Riddell CA, Kaufman JS. Beyond intention to treat: what is the right question? *Clin Trials* 2014;11:28-37.

(11) Hernan MA, Hernandez-Diaz S. Beyond the intention-to-treat in comparative effectiveness research. *Clin Trials* 2012;9:48-55.

(12) Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf* 2010;19:537-554.

(13) Toh S, Hernan MA. Causal inference from longitudinal studies with baseline randomization. *Int J Biostat* 2008;4:Article.

(14) Tilbrook HE, Hewitt CE, Aplin JD et al. Compliance effects in a randomised controlled trial of yoga for chronic low back pain: a methodological study. *Physiotherapy* 2014;100:256-262.

(15) Van Lieshout J, Felix-Schollaart B, Bolsius EJM. NHG Standaard Schildklieraandoeningen (tweede herziening). *Huisarts Wet* 2013;56:320-330.

(16) Swanson SA, Miller M, Robins JM, Hernan MA. Definition and Evaluation of the Monotonicity Condition for Preference-based Instruments. *Epidemiology* 2015.

(17) Swanson SA, Hernan MA. Commentary: how to report instrumental variable analyses (suggestions welcome). *Epidemiology* 2013;24:370-374.

(18) Angrist JD, Pischke JS. *Mostly Harmless Econometrics: An Empiricist's Companion*. 2008.

(19) Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006;17:360-372.

(20) Balke A, Pearl J. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* 1997;92:1171-1176.

(21) Wehby GL, Ohsfeldt RL, Murray JC. 'Mendelian randomization' equals instrumental variable analysis with genetic instruments. *Stat Med* 2008;27:2745-2749.
(22) Lawlor DA, Windmeijer F, Smith GD. Is Mendelian randomization 'lost in translation?': comments on 'Mendelian randomization equals instrumental variable analysis with genetic instruments' by Wehby et al. *Stat Med* 2008;27:2750-2755.
(23) Vanderweele TJ, Tchetgen Tchetgen EJ, Cornelis M, Kraft P. Methodological challenges in mendelian randomization. *Epidemiology* 2014;25:427-435.

Nederlandse samenvatting

Introductie

Voor het beoordelen van een effect van een behandeling worden resultaten van een gerandomiseerde trial vaak als ideaal beschouwd. Gegevens uit gerandomiseerde trials zijn voor veel belangrijke vragen rondom medische behandelingen echter niet beschikbaar. Er is daarom dringend behoefte aan methoden waarbij de effecten van een behandeling op valide wijze geschat kunnen worden in observationele data. Een belangrijk probleem in observationeel onderzoek is dat patiëntkarakteristieken meewegen bij behandelingsbeslissingen: behandelde patiënten en niet-behandelde patiënten (of patiënten behandeld met A en patiënten behandeld met B) zijn daardoor niet vergelijkbaar. Anders geformuleerd, factoren (kenmerken van de patiënt) die zowel de behandelingsbeslissing als de uitkomst beïnvloeden zullen het effect van de behandeling vertekenen. Dit wordt *confounding by indication* genoemd. Gebruikelijke analysemethodes kunnen alleen corrigeren voor die kenmerken van de patiënt die gemeten zijn en nemen daarbij aan dat er geen verdere ongemeten factoren zijn die het effect nog vertekenen, wat veelal niet plausibel is.

Een methode, afkomstig uit de econometrie, die in potentie zowel gemeten als ongemeten confounding kan omzeilen is de instrumentele variabele analyse. Deze methode zoekt een surrogaat voor de randomisatie; een factor die de behandeling mede bepaalt, maar niet op een andere manier dan via de behandeling geassocieerd is met de uitkomst. Voorbeelden zijn afstand tot een ziekenhuis met bepaalde behandelingsmogelijkheden, of behandelingsvoorkeur van ofwel het ziekenhuis, ofwel de arts. In formelere zin moet een instrumentele variabele (instrument) aan 3 hoofdaannames voldoen: 1) het instrument is geassocieerd met de behandeling; 2) het instrument heeft geen effect op de uitkomst, anders dan via de behandeling; 3) er zijn geen factoren die een effect hebben op zowel het instrument als op de uitkomst. Als aan deze aannames wordt voldaan is een associatie tussen het instrument en de uitkomst volledig toe te schrijven aan verschillen in de behandeling. Hiermee kan het oorzakelijke effect van de behandeling op de uitkomst geschat worden.

Het doel van dit proefschrift was te onderzoeken hoe valide instrumentele variabele analyse is als methode om zowel beoogde effecten als bijwerkingen van een behandeling te bepalen, en voor welke vragen en in welke situaties deze analyse bruikbaar is.

Voorschrijfvoorkeur als instrumentele variabele

We hebben ons vooral gericht op het gebruik van 'voorschrijfvoorkeur' als instrumentele variabele. De keuze voor een behandeling door artsen wordt bepaald

door zowel kenmerken van de patiënt als door een onderliggende voorkeur van de arts voor een bepaalde behandeling. Door verschillen in deze onderliggende voorkeur kan het zijn dat verschillende artsen verschillende behandelingskeuzes zouden maken voor dezelfde patiënt. Door verschillen in onderliggende voorkeur ontstaat dus variatie in behandeling die onafhankelijk is van patiëntkenmerken: ofwel, onderliggende behandelingsvoorkeur (of 'voorschrijfvoorkeur') kan fungeren als instrumentele variabele. Het concept 'voorschrijfvoorkeur' hebben we nader onderzocht in **Hoofdstuk 2**: bestaat er echt variatie in voorkeur voor behandeling, of komt de variatie in keuze voor behandeling tussen artsen door verschillen in hun patiëntenpopulatie? We bekeken gegevens uit een enquête onder huisartsen, waarin werd gevraagd of zij acht fictieve patiënten met subklinische hypothyreoïdie zouden behandelen met levothyroxine. We vonden aanzienlijke variatie in de keuze om patiënten al dan niet met levothyroxine te behandelen, waaruit we concluderen dat verschillen in voorschrijfvoorkeur bestaan.

Om met een instrumentele variabele analyse een puntschatting te krijgen van het effect van de behandeling is naast de hierboven besproken 3 hoofdaannames nog een vierde aanname nodig. Er bestaan verschillende mogelijkheden voor deze vierde aanname, en voor elk van deze mogelijkheden is de interpretatie van het geschatte effect iets anders. In **Hoofdstuk 2** concluderen we op basis van de voorschrijfpatronen uit de enquête dat voor voorschrijfvoorkeur de zgn. stochastische monotoniciteitsassumptie aannemelijk is (maar de strengere deterministische monotoniciteitsassumptie niet). De interpretatie van de effectschatting onder deze aanname is ingewikkeld: het is een gewogen gemiddelde van de effecten in bepaalde subgroepen van de studiepopulatie. Hoe sterker het instrument gerelateerd is aan de behandeling in de subgroep, hoe zwaarder het effect in die subgroep meeweegt.

In welke situaties is instrumentele variabele analyse een geschikte analyse?
In **Hoofdstuk 3** werd instrumentele variabele analyse met voorschrijfvoorkeur als instrument toegepast in een studie met enkele honderden patiënten, een populatiegrootte die gebruikelijk is in epidemiologische studies. Schattingen van het effect van preoperatieve corticosteroïden op beademingsduur, verblijf op de intensive care en in het ziekenhuis, het optreden van infecties, boezemfibrilleren, hartfalen en delier bij patiënten die een hartoperatie ondergingen waren qua richting hetzelfde als effectschattingen uit een gerandomiseerde trial. De gevonden effecten waren echter veel groter, en de schattingen waren zo onnauwkeurig dat ze weinig informatief waren. Hieruit concluderen we dat de geringe precisie van instrumentele variabele analyse een grote beperking vormt voor de toepasbaarheid in studies met patiëntaantallen die voor gebruikelijke analyses groot genoeg zijn.

Als er sprake is van ongemeten confounding zal een analyse waarbij alleen gecorrigeerd wordt voor gemeten confounding een systematisch vertekende schatting geven: er is sprake van *bias*. Als een instrumentele variabele beschikbaar is die aan de drie hoofdaannames voldoet dan zal de schatting uit de instrumentele variabele analyse niet systematisch vertekend zijn (enkele kanttekeningen hierbij laten we hier buiten beschouwing), maar deze schatting is veel onnauwkeuriger: de variantie is veel groter. Welke schatting gemiddeld dichter bij het ware effect zal zijn is een *trade-off* tussen de bias van de gebruikelijke analyse en de variantie van de instrumentele variabele analyse. In **Hoofdstuk 4** hebben we in een simulatiestudie laten zien dat de richting waarin deze balans uitslaat in grote mate afhangt van de grootte van de studiepopulatie, en daarnaast van de sterkte van het instrument en de sterkte van de ongemeten confounding. We hebben een formule opgesteld waarmee kan worden berekend vanaf welke studiepopulatiegrootte de instrumentele variabele schatting gemiddeld minder ver afwijkt van het ware effect dan de gebruikelijke schatting. Dit geldt in het algemeen pas vanaf zeer grote aantallen .

In **Hoofdstuk 5** werd onderzocht of in observationele studies naar bijwerkingen instrumentele variabele analyse bruikbaar is als sensitiviteitsanalyse om te beoordelen of er sprake is van confounding. In een grote groep nieuwe gebruiksters van gecombineerde hormonale anticonceptiepillen hebben we het risico op diep veneuze trombose en longembolieën vergeleken tussen gebruikers van de tweede en derde generatiepil. In principe is dit een weinig voorspelbare bijwerking: we verwachtten daarom niet dat de soort pil die wordt voorgeschreven gerelateerd is aan het onderliggende risico op trombose van de patiënt (geen confounding). Als dit het geval is zou een instrumentele variabele analyse een zelfde schatting geven als een gebruikelijke analyse. Dit is ook wat we vonden: de schattingen uit beide soorten analyses suggereerden een hoger risico op trombose voor de derde generatiepil. Echter, zelfs in deze zeer grote studiepopulatie van enkele honderdduizenden vrouwen was de instrumentele variabele schatting onnauwkeurig, omdat trombose een zeldzame uitkomst is. Vanwege de overeenkomst tussen de resultaten van de verschillende analyses die rusten op verschillende aannames concluderen we dat de aanwezigheid van een aanzienlijke hoeveelheid confounding onwaarschijnlijk is.

Het rapporteren van een instrumentele variabele analyse

In **Hoofdstuk 6** gaven we de aanbeveling om in een studie waarin een instrumentele variabele analyse wordt gedaan altijd eerst te laten zien wat de associatie is tussen de instrumentele variabele en de uitkomst (voordat de formele instrumentele variabele analyse wordt gedaan). Dit is belangrijk, omdat dit de associatie is die in een formele instrumentele variabele analyse wordt geëxtrapoleerd. Deze aanbeveling hebben

we gedaan in reactie op een artikel waarin richtlijnen voor het rapporteren van een instrumentele variabele analyse waren opgesteld.

Mendeliaanse randomisatie

In de laatste twee hoofdstukken hebben we ons gericht op een andere vorm van instrumentele variabele analyse: mendeliaanse randomisatie. Hierbij worden genetische varianten als instrumentele variabele gebruikt om het effect van een blootstelling op een uitkomst te onderzoeken. Het idee is dat genetische varianten willekeurig overerven van ouders naar hun kinderen en dat deze varianten daarom over het algemeen niet gerelateerd zijn aan andere factoren die een effect hebben op de uitkomst (zoals leefstijlfactoren). Het effect op de uitkomst van deze genetische variatie in de blootstelling zou daarom niet vertekend moeten worden door deze andere factoren. In **Hoofdstuk 7** hebben we d.m.v. een systematische review onderzocht wat voor aanpak in mendeliaanse randomisatie studies werd gebruikt. Hierbij werden drie hoofdcategorieën onderscheiden: 1) alleen onderzoeken van de associatie tussen het genetische instrument en de uitkomst; 2) een formele instrumentele variabele analyse met een genetisch instrument; 3) een vergelijking van de geobserveerde en de verwachte associatie tussen het genetische instrument en de uitkomst. Verder hebben we gevonden dat de hoofdaannames voor een instrumentele variabele, die ook gelden voor Mendeliaanse randomisatie, vaak onvoldoende besproken werden en dat de gebruikte statistische methoden vaak niet duidelijk gerapporteerd werden. Daarom hebben we een checklist opgesteld voor het rapporteren van mendeliaanse randomisatie studies.

In **Hoofdstuk 8** hebben we uitgelegd dat in mendeliaanse randomisatiestudies in oudere populaties (bijv. >80 jaar) sprake kan zijn van vertekening van het onderzochte effect door de selectie op leeftijd. De genetische variant die de instrumentele variabele is in een dergelijke studie, kan invloed hebben gehad op overleving tot 80 jaar. Andere factoren, zoals roken, zullen de overleving tot 80 jaar beïnvloed hebben. Hoewel op jonge leeftijd geen verband zal bestaan tussen de genetische variant en rookgedrag, kan dit verband er in een oudere populatie wel zijn: extreem geformuleerd zullen mensen die ondanks een ongunstige genetische variant de leeftijd van 80 jaar hebben bereikt waarschijnlijk niet roken. Als dit verband bestaat voldoet de genetische variant niet meer aan één van de onderliggende aannames van mendeliaanse randomisatie: aan de aanname dat het instrument niet gerelateerd is aan andere factoren die een effect hebben op de uitkomst wordt niet voldaan.

Conclusies

Een belangrijke conclusie van dit proefschrift is dat instrumentele variabele analyse alleen in zeer grote epidemiologische studies met aanzienlijke ongemeten confounding en een sterk verband tussen het instrument en de behandeling of blootstelling zinvol is als primaire analyse. Veel epidemiologische studies hebben een studiepopulatie van enkele honderden personen. In dit geval zal een schatting uit een instrumentele variabele analyse dusdanig onnauwkeurig zijn dat het weinig informatie geeft. Een instrumentele variabele analyse zal wellicht vaker geschikt zijn als sensitiviteitsanalyse, bijvoorbeeld om te beoordelen of er sprake is van confounding in studies naar bijwerkingen.

Verder hebben we gevonden dat artsen verschillen in hun behandelingsvoorkeur voor eenzelfde patiënt, onafhankelijk van kenmerken van deze artsen en hun patiëntpopulaties. De stochastische monotoniciteitsassumptie lijkt een plausibele aanname om een puntschatting te verkrijgen, (als het behandelingseffect niet gelijk verondersteld kan worden). Een probleem hierbij is dat het geschatte effect uit een IV analyse moeilijk interpreteerbaar is.

Tenslotte concluderen we dat door mendeliaanse randomisatiestudies als instrumentele variabele studies te beschouwen het rapporteren van mendeliaanse randomisatiestudies kan worden verbeterd, ook als geen formele instrumentele variabele analyse wordt verricht.

Acknowledgements

Several of the projects in this thesis were in collaboration with other departments: with the Department of Intensive Care of the LUMC, with the Department of Public Health and Primary Care of the LUMC (and all others involved in the general practitioner survey project) and with the Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University. I would hereby like to thank all those who were involved in these fruitful collaborations. I would also like to thank all colleagues in the department of Clinical Epidemiology for their guidance, advice, discussions and the essential coffee breaks. Lastly I would like to thank my family and friends for their continuous support.

Publications

Boef AGC, le Cessie S, Dekkers OM, Frey P, Kearney PM, Kerse N, Mallen CD, McCarthy VJC, Mooijaart SP, Muth C, Rodondi N, Rosemann T, Russell A, Schers H, Virgini V, de Waal MWM, Warner A, Gussekloo J, den Elzen WPJ. Physician's prescribing preference as an instrumental variable: exploring assumptions using survey data. *Epidemiology.* 2015 Nov 24. (Epub ahead of print)

Boef AGC, Dekkers OM, le Cessie S. Mendelian randomization studies: a review of the approaches used and the quality of reporting. *Int J Epidemiol.* 2015; 44(2):496-511.

Boef AGC, le Cessie S, Dekkers OM. Mendelian randomization studies in the elderly. *Epidemiology.* 2015; 26(2):e15-6.

Boef AGC, Dekkers OM, Vandenbroucke JP, le Cessie S. Sample size importantly limit the usefulness of instrumental variable analysis, depending on confounding and instrument strength. *J Clin Epidemiol*. 2014; 67:1258-1264.

Boef AGC*, van Paassen J*, Arbous MS, Middelkoop A, Vandenbroucke JP, le Cessie S, Dekkers OM. Physician's preference-based instrumental variable analysis: is it valid and useful in a moderate-sized study? *Epidemiology.* 2014; 25(6):923-7.

Boef AGC, Postmus I, Siegerink B. Re: "Mendelian randomization and estimation of treatment efficacy for chronic diseases". *Am J Epidemiol.* 2014; 179(2):264.

Boef AGC, Dekkers OM, le Cessie S, Vandenbroucke JP. Reporting instrumental variable analyses. *Epidemiology.* 2013; 24(6):937-8.

Eriksson UK, van Bodegom D, May L, Boef AGC, Westendorp RGJ. Low C-reactive protein levels in a traditional West-African population living in a malaria endemic area. *PLoS One.* 2013; 8(7):e70076.

Boef AGC, le Cessie S, Dekkers OM. Instrumentele-variabele-analyse. *Ned Tijdschr Geneeskd.* 2013;157(4):A5481.

Boef AGC, May L, van Bodegom D, van Lieshout L, Verweij JJ, Maier AB, Westendorp RGJ, Eriksson UK. Parasitic infections and immune function: effect of helminth infections in a malaria endemic area. *Immunobiology.* 2013; 218(5): 706-711.

Boef AGC, May L, van Bodegom D, Kuningas M, Eriksson UK, Westendorp RGJ. The influence of genetic variation on innate immune activation in an environment with high infectious pressure. *Genes Immun.* 2012; 13(2):103-8.

Curriculum Vitae

Anna Gunnel Christina Boef werd geboren op 18 september 1988 in Zoeterwoude. Na vier jaar aan Cults Academy/Rijnlands Lyceum Aberdeen in Cults, Aberdeen, Verenigd Koninkrijk en twee jaar aan het Penta College CSG Jacob van Liesveldt te Hellevoetsluis behaalde zij in 2005 haar gymnasium diploma. Aansluitend begon zij aan haar studie Geneeskunde aan het Leids Universitair Medisch Centrum. In 2011 behaalde zij haar doctoraalexamen (cum laude) en haar artsexamen. Van januari 2012 tot januari 2015 deed zij promotieonderzoek aan de afdeling Klinische Epidemiologie onder begeleiding van prof. dr. Jan Vandenbroucke, dr. Saskia le Cessie en dr. Olaf Dekkers, wat heeft geleid tot dit proefschrift. Van april tot en met december 2015 was ze werkzaam als ANIOS interne geneeskunde in het Medisch Centrum Haaglanden.

Stellingen behorend bij het proefschrift

Obtaining causal estimates of therapeutic effects in observational studies:

the usefulness and validity of
physician's preference as an instrumental variable.

1. Instrumental variable analysis is primarily useful as a complementary analysis rather than as a primary analysis within clinical epidemiology. *(this thesis)*

2. The population in which an instrumental variable study is performed will rarely be the population to which the instrumental variable estimate applies. *(this thesis)*

3. Although variation in treatment preference among physicians exists for many therapeutic questions, the potential for physician's prescribing preference to be used as an instrumental variable is limited because it is not a directly measurable characteristic. *(this thesis)*

4. Mendelian randomisation studies should be viewed as and reported as an instrumental variable study, even if no formal instrumental variable analysis is performed. *(this thesis)*

5. In a study in which considerable confounding can be expected, one should be aware that the existence of a very strong instrument within the IV assumptions is impossible.
E.P. Martens, W.R. Pestman, A. de Boer, S.V. Belitser, O.H. Klungel, Epidemiology 2006;17(3): 260-7.

6. IV methods are not an epidemiologist's dream come true.
M.A. Hernán and J.M. Robins, Epidemiology 2006;17(4):360-72.

7. Even if a doctor's every fully articulated thought regarding a treatment decision could be recorded and adjusted for, confounding by indication in observational studies of treatment effects would still not be completely resolved.

8. Propensity score methods have a high propensity for being misinterpreted.

9. The peer review process would benefit from reviewers being reviewed.

Anna G.C. Boef
Leiden, 10 februari 2016