Journal of Health Economics 45 (2016) 131–148

Contents lists available at ScienceDirect

Journal of Health Economics

journal homepage: www.elsevier.com/locate/econbase

# Genetic markers as instrumental variables

Stephanie von Hinke[a,*,1], George Davey Smith[a], Debbie A. Lawlor[a], Carol Propper[a,b], Frank Windmeijer[a]

[a] University of Bristol, Bristol, United Kingdom
[b] Imperial College London, London, United Kingdom

ARTICLE INFO

*Article history:*
Received 4 September 2014
Received in revised form 7 August 2015
Accepted 19 October 2015
Available online 10 November 2015

*Keywords:*
ALSPAC
Genetic variants
Instrumental variables
Mendelian randomization
Potential outcomes

*JEL codes:*
C36, I1, I2

ABSTRACT

The use of genetic markers as instrumental variables (IV) is receiving increasing attention from economists, statisticians, epidemiologists and social scientists. Although IV is commonly used in economics, the appropriate conditions for the use of genetic variants as instruments have not been well defined. The increasing availability of biomedical data, however, makes understanding of these conditions crucial to the successful use of genotypes as instruments. We combine the econometric IV literature with that from genetic epidemiology, and discuss the biological conditions and IV assumptions within the statistical potential outcomes framework. We review this in the context of two illustrative applications.



## 1. Introduction

Many studies in the social and epidemiological sciences aim to make causal inferences using observational data. This is often problematic, as observed associations are not necessarily causal, with confounding being an important concern. Randomization of treatment, as in a Randomized Controlled Trial (RCT), is one way to infer causality. However, such experiments are not always possible or feasible. An approach commonly used in the economics and econometrics literature is that of Instrumental Variables (IV). This introduces a third variable (the instrument) that is robustly associated with the risk factor of interest, but not with the outcome variable, other than through its effect on the risk factor. This instrument can then be exploited to make causal inferences about the effect of the risk factor on the outcome.

Recently, epidemiologists, statisticians, economists and other social scientists have become interested in using genetic variants as instruments. 'Mendelian randomization' refers to the random assignment of an individual's genotype at conception (Davey Smith and Ebrahim, 2003; Davey Smith, 2007). Under certain assumptions that we discuss in detail below, observed associations between genetic variants and the outcome of interest are unlikely to be due to confounding by behavioural or environmental factors. Mendelian randomization can therefore be exploited to make causal inferences about the effects of modifiable (non-genetic) risk factors, on different outcomes[1]. Statisticians have highlighted some of the implicit statistical assumptions commonly made in Mendelian randomization studies (e.g. Didelez and Sheehan, 2007; Didelez et al., 2010). Genetic epidemiologists emphasize the importance of carefully examining the conditions that need to be met for genetic variants to be used as instruments (see e.g. Davey Smith and Ebrahim, 2003; Sheehan et al., 2008; Lawlor et al., 2008a,b). However, while studies in economics commonly use IV methods, the (biological) conditions relevant for Mendelian randomization have not been disseminated widely in this literature. The increasing availability of biomedical information in social science

* Corresponding author. Tel.: +44 1173317936.
*E-mail address:* S.vonHinke@bristol.ac.uk (S. von Hinke).

[1] We thank the editor Adriana Lleras-Muney, two anonymous reviewers, John Cawley, Paul Clarke, Neil Davies, George Leckie, Owen O'Donnell, Jennifer Roberts, Sarah Smith, Nic Timpson, Gerard van den Berg, and George Wehby for helpful suggestions. We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The views expressed in this paper are those of the authors and not necessarily any people mentioned in these acknowledgements.

[1] Appendix A provides a brief guide to the terms used in genetic studies.

http://dx.doi.org/10.1016/j.jhealeco.2015.10.007

datasets, however, makes understanding them crucial to the successful use of genotypes as instruments for modifiable risk factors.

The contribution of this paper is to discuss these conditions within the well-known statistical potential outcomes framework. We use the work by Imbens and Angrist (1994), Angrist et al. (1996), and Angrist et al. (2000), among others, which has been of great importance in linking the econometric IV literature to the potential outcomes framework. We link Mendelian randomization to this framework, and discuss how the conditions, defined in genetic epidemiology, relate to the IV assumptions used in statistics and economics. To communicate best practice in genetic epidemiology to a wider economics audience, we review these conditions in the context of two illustrative applications: one in social science and one in medicine. Specifically, we examine whether child fat masscausally affects (1) academic achievement, and (2) blood pressure, using 32 recently identified genetic variants as instrumental variables for fat mass.

These examples are pertinent for several reasons. For our social science application for example, obese children are more likely to be absent from school, have sleep disorders, and be treated differently by teachers, parents and peers. All these may affect children's (learning) environment and educational outcomes. However, an observed association between fat mass and academic achievement is not necessarily causal. There are likely to be many confounders, and one can never be sure that all relevant ones are accounted for. For our medical application, there is evidence that even relatively small reductions in weight can reduce blood pressure and hypertension risk (Neter et al., 2003). However, the increase in obesity in recent decades has been accompanied by a decrease in hypertension, leading to questions about their association, with some suggesting that randomized controlled trials of weight reduction could have affected blood pressure through mechanisms other than weight loss (Campos et al., 2006). The use of ordinary least squares (OLS) suggests that fat mass is inversely related to educational attainment, but increases the risk of hypertension. When using carefully selected genetic variants as instruments for fat mass, we find no evidence of a causal relationship between fat mass and academic performance, although the parameters are imprecisely estimated. In contrast, we find a positive effect of fat mass on blood pressure, suggesting that reductions in fat mass will reduce the risk of cardiovascular disease.

Although Mendelian randomization is widely used in the medical and epidemiological sciences, with its findings being fed into pharmacotherapeutic development, it is very controversial within economics. This mainly stems from the credibility of the 'exclusion restriction': the assumption that the variants do not directly affect the outcome of interest. Indeed, there are many situations that may violate this assumption, invalidating the instruments and biasing the estimates. One of the issues is that we have very limited knowledge and understanding of the specific functions of genes, and studies that directly examine gene-function are often underpowered. Hence, we can never be certain that the exclusion restriction is satisfied. We discuss this in detail, and highlight the specific (biological) pathways through which the use of genetic variants as instrumental variables may lead to invalid inferences, including the potential for variants to have multiple functions, or to be correlated to other variants that affect the outcome of interest. We also consider the implications of gene–gene and gene–environment interactions for Mendelian randomization. Finally, it is worth noting that the uncertainty of the exclusion restriction is not specific to Mendelian randomization. Indeed, any IV analysis relies on this untestable assumption, and one generally assesses such studies based on whether the available evidence suggests that the assumption is likely to hold (see also von Hinke et al., 2012). We discuss different ways of exploring its validity indirectly in the context of Mendelian randomization and attempt to clearly articulate the potential situations that would invalidate the approach[2].

Section 2 details the conditions that need to be met for genetic variants to be used as instruments. Section 3 introduces our empirical application. We describe the data, examine the validity of our choice of genetic variants, present the results as well asa number of sensitivity checks. Section 4 concludes and discusses the implications of our findings in terms of best practice for using Mendelian randomization by researchers who do not come from a primarily biological discipline.

## 2. The use of genetic variants as instrumental variables

We start by discussing the links between Mendelian randomization and other approaches used in the medical and social science literature. We then build on the Potential Outcomes Framework by Imbens and Angrist (1994), Angrist and Imbens (1995), Angrist et al. (1996, 2000). We first briefly outline the well-known structural assumptions in the context of our applications, and then discuss how Mendelian randomization links to the statistical assumptions of this framework.

### 2.1. Mendelian randomization

We discuss Mendelian randomization from a statistics and economics perspective in the context of a social study, with the aim of making causal inferences about the effect of a treatment on an outcome of interest. Depending on the discipline, the terms 'treatment', 'risk factor', 'exposure', 'predictor', or 'intermediate phenotype' have all been used to denote the variable of interest that potentially causes the outcome. To avoid confusion, the remainder of this paper uses either the term 'treatment' or 'risk factor'.

The concept of Mendelian randomization is closely linked to Randomized Controlled Trials (RCTs), where the allocation of treatment is randomized over all eligible individuals (Davey Smith and Ebrahim, 2005; Hingorani and Humphries, 2005). Indeed, IV can be applied to analyse encouragement designs (such as RCTs where the instrument is the encouragement to participate) that are affected by non-ignorable non-compliance. Non-compliance refers to the fact that individuals can choose to take or not take treatments other than those to which they are randomized. Non-ignorable non-compliance refers to participants choosing to take or not to take the treatment that they are randomized to in a manner associated with their study outcomes, after adjusting for baseline characteristics. This is also known as endogenous treatment in economics, or selection into treatment.

The idea is similar for the social context in our application: individuals 'select' their treatment – fat mass – through lifestyle choices, such as diet and physical activity, which are likely to be related to their study outcome (educational attainment and blood pressure). In a well-conducted RCT of an intervention aimed at reducing fat mass, the random allocation effectively balances these lifestyle choices between groups. Comparing groups based on the original random allocation ('intention to treat') maintains this balance, whereas comparing groups based on what treatment was actually chosen by the participant (a 'per-protocol' analysis) is likely to be biased due to non-ignorable non-compliance. In other words, treatment by choice (as opposed to treatment by randomization) is likely to be related to the outcome through characteristics such as social class, income, diet, etc.

There are many cases, however, where RCTs are infeasible (for example, there may be no effective intervention to randomize, such

[2] Thus we give below examples of situations where the use of genetic variants as instrumental variables is *more* as well as *less* likely to lead to incorrect inference.