Economics and Human Biology 13 (2014) 99–106

Contents lists available at ScienceDirect

Economics and Human Biology

journal homepage: http://www.elsevier.com/locate/ehb

Mendelian randomization in health research: Using appropriate genetic variants and avoiding biased estimates

Amy E. Taylor [a,*], Neil M. Davies [b], Jennifer J. Ware [a,b], Tyler VanderWeele [c,d], George Davey Smith [b], Marcus R. Munafò [a]

[a] MRC Integrative Epidemiology Unit (IEU) at the University of Bristol, UK Centre for Tobacco and Alcohol Studies, School of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1TU, UK
[b] MRC Integrative Epidemiology Unit (IEU) at the University of Bristol, School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, UK
[c] Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA
[d] Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

ARTICLE INFO

Article history:
Received 15 May 2013
Received in revised form 3 December 2013
Accepted 3 December 2013
Available online 13 December 2013

Keywords:
Smoking
Tobacco
Mendelian randomization
Causal inference
Instrumental variable

ABSTRACT

Mendelian randomization methods, which use genetic variants as instrumental variables for exposures of interest to overcome problems of confounding and reverse causality, are becoming widespread for assessing causal relationships in epidemiological studies. The main purpose of this paper is to demonstrate how results can be biased if researchers select genetic variants on the basis of their association with the exposure in their own dataset, as often happens in candidate gene analyses. This can lead to estimates that indicate apparent "causal" relationships, despite there being no true effect of the exposure. In addition, we discuss the potential bias in estimates of magnitudes of effect from Mendelian randomization analyses when the measured exposure is a poor proxy for the true underlying exposure. We illustrate these points with specific reference to tobacco research.

1. Introduction

Proving how exposures affect health outcomes can be problematic in observational studies. Even if an exposure and an outcome are associated, the direction of causality can be difficult to ascertain because health outcomes can lead to changes in behaviour which can affect exposures (Munafò and Araya, 2010). Mendelian randomization studies may help to shed light on these relationships by using genetic variants, such as single nucleotide polymorphisms (SNPs) (see Table 1 for definition), as instrumental variables for measured lifestyle exposures (Davey Smith and Ebrahim, 2003). Mendelian randomization studies can be used for two related purposes: (1) to provide evidence for the *existence* of causal associations, and (2) to enable *accurate estimation* of the magnitude of the effect of lifelong exposure to a risk factor on an outcome (Davey Smith and Ebrahim, 2004).

As is the case for instrumental variable methods generally, for Mendelian randomization studies to be useful genetic variants must be robustly associated with the exposure of interest (Davey Smith and Ebrahim, 2005; Lawlor et al., 2008b). Despite this, recent Mendelian randomization studies conducted by Wehby et al. (2011a,b, 2012) have used genetic variants as instruments for smoking heaviness which were not shown to be

* Corresponding author.
Tel.: +44 0117 3318239; fax: +44 0117 9288588.
E-mail address: amy.taylor@bristol.ac.uk (A.E. Taylor).

1570-677X
http://dx.doi.org/10.1016/j.ehb.2013.12.002

Table 1
Definitions of genetic terms for Mendelian randomization.

Term	Definition
Allele	One form of a genetic variant
Canalization	Process of developmental compensation for the effects of a genetic variant which may disrupt normal development
Genetic variant	Part of the genetic code for which there is more than one form in the population. This can be a single nucleotide polymorphism but other forms of variation exist
Genome wide association study (GWAS)	Hypothesis-free study which investigates associations of a large number of genetic variants across the whole genome with a trait of interest
Linkage disequilibrium	Non-random association between genetic variants at different positions along the chromosome
Pleiotropic	Influencing more than one phenotypic trait
Single nucleotide polymorphism (SNP)	Variation at a single nucleotide base pair in the DNA sequence

associated with smoking phenotypes in large genome wide association studies. Whilst the authors acknowledge that these variants have not been consistently associated with smoking phenotypes, they suggest that the variants provide evidence of causal effects of smoking on body weight (Wehby et al., 2012) and smoking in pregnancy on birthweight (Wehby et al., 2011b) and risk of orofacial clefts in offspring (Wehby et al., 2011a). In addition, the authors use the genetic variants to estimate the magnitude of effect of smoking heaviness on their outcomes of interest (Wehby et al., 2011a,b, 2012). Even if the variants they use are truly associated with smoking behaviour, this is likely to produce incorrect estimates of the effect size of smoking on the outcome.

1.1. Aims

In this paper, we aim: (1) to illustrate, using a data simulation, why inferences based on the results of Mendelian randomization studies using genetic variants selected based on their association in a single sample are likely to be misleading and (2) to demonstrate why estimating the magnitudes of causal effects in cases where the measured exposure is not the same as the underlying exposure captured by the variant is problematic. We discuss these issues with reference to the specific case of tobacco as an exposure, but these principles can be applied more widely to Mendelian randomization and instrumental variable analyses.

1.2. Assumptions of Mendelian randomization

The principle of Mendelian randomization relies on the basic (but approximate) laws of Mendelian genetics (segregation and independent assortment). If these two laws hold, then at a population level, genetic variants will not be associated with the confounding factors that generally distort conventional observational studies (Davey Smith and Ebrahim, 2003; Davey Smith, 2011). In addition, genetic variants will not be affected by reverse causality (Davey Smith and Ebrahim, 2003). Epidemiological studies increasingly use Mendelian randomization to provide robust evidence of underlying causal mechanisms in a number of areas of health research including cardiovascular disease, cancer and mental health (Casas et al., 2005; Davey Smith et al., 2005; Benn et al., 2011; Scott et al., 2011; Interleukin-6 Receptor Mendelian Randomisation Analysis et al., 2012; Nordestgaard et al., 2012; Voight et al., 2012; Carslake et al., 2013).

For a SNP to be a valid instrumental variable, the following assumptions must hold: (1) the SNP should be reliably associated with the exposure, (2) the SNP should only be associated with the outcome through the exposure of interest (the "exclusion restriction") and (3) the SNP should be independent of other factors affecting the outcome (confounders) (Angrist et al., 1996; Lawlor et al., 2008b; Wehby et al., 2008; Clarke and Windmeijer, 2012). Moreover, to use Mendelian randomization for accurate estimation of effect sizes in mediation analysis using a measured exposure, the measured exposure should accurately capture the true causal exposure (Lawlor et al., 2008a; Pierce and VanderWeele, 2012).

2. Use of genetic variants selected in a single sample

2.1. Genetic variants for tobacco research

Large consortium-based genome wide association studies have found genetic variants robustly associated with smoking behaviours (Thorgeirsson et al., 2008; Furberg et al., 2010; Liu et al., 2010). One genetic variant that has been highlighted by these studies, amongst others, is located in the nicotinic receptor gene cluster *CHRNA5–A3–B4* on chromosome 15. Two SNPs within this region, rs16969968 and rs1051730, which are in linkage disequilibrium and can be used interchangeably in studies on Europeans, consistently associate with measures of heaviness of smoking (e.g., cigarettes per day or biomarkers of nicotine exposure) (Freathy et al., 2009; Munafò et al., 2012). Smokers with a single copy of the smoking increasing allele smoke on average one extra cigarette per day compared to those with no copies. The effects of the SNP are additive, so people with two copies of the smoking increasing allele on average smoke two additional cigarettes a day (Ware et al., 2011). The strength and consistency of this association make these variants suitable instruments for use in Mendelian randomization studies. The second assumption of instrumental variable analysis, that the SNP should only be associated with the outcome through the exposure of interest, is rarely fully testable (Glymour et al., 2012). In Mendelian randomization, this assumption may be violated if the genetic variant has pleiotropic effects, is in linkage disequilibrium with another variant of differing function or if its effects are