THE CENTRE FOR MARKET AND PUBLIC ORGANISATION

Genetic Markers as Instrumental Variables

Stephanie von Hinke Kessler Scholder, George Davey Smith, Debbie A. Lawlor, Carol Propper and Frank Windmeijer

October 2011

Working Paper No. 11/274

Centre for Market and Public Organisation
Bristol Institute of Public Affairs
University of Bristol
2 Priory Road
Bristol BS8 1TX
http://www.bristol.ac.uk/cmpo/

Tel: (0117) 33 10952
Fax: (0117) 33 10705
E-mail: cmpo-admin@bristol.ac.uk

ISSN 1473-625X

CMPO Working Paper Series No. 11/274

Genetic Markers as Instrumental Variables

Stephanie von Hinke Kessler Scholder[1], George Davey Smith[2],
Debbie A. Lawlor[2], Carol Propper[3], Frank Windmeijer[4]

[1]*University of York*
[2]*University of Bristol*
[3] *CMPO, University of Bristol and Imperial College London*
[4]*CMPO, University of Bristol*

October 2011

Abstract
The use of genetic markers as instrumental variables (IV) is receiving increasing attention from epidemiologists, economists, statisticians and social scientists. This paper examines the conditions that need to be met for genetic variants to be used as instruments. Although these have been discussed in the epidemiological, medical and statistical literature, they have not been well-defined in the economics and social science literature. The increasing availability of biomedical data however, makes understanding of these conditions crucial to the successful use of genotypes as instruments for modifiable risk factors. We combine the econometric IV literature with that from genetic epidemiology using a potential outcomes framework and review the IV conditions in the context of a social science application, examining the effect of child fat mass on academic performance.

Keywords: ALSPAC; Fat mass; Genetic Variants; Instrumental Variables; Mendelian Randomization; Potential Outcomes

JEL Classification: C36, I1, J24

Electronic version: www.bristol.ac.uk/cmpo/publications/papers/2011/wp274.pdf

Acknowledgements
We thank the editor, the associate editor, anonymous reviewers, John Cawley, Paul Clarke, George Leckie, Owen O'Donnell, Jennifer Roberts, Sarah Smith, Nicholas 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.

Address for correspondence
CMPO, Bristol Institute of Public Affairs
University of Bristol
2 Priory Road
Bristol BS8 1TX
cmpo-admin@bristol.ac.uk
www.bristol.ac.uk/cmpo/

1. Introduction

Many studies in the social and epidemiological sciences aim to make causal inferences using observational data. This is often problematic, as it is not always clear which of any two associated variables is the cause and which is the effect, or whether other unobserved factors affect both variables. Randomization of treatment, like that in a Randomized Controlled Trial (RCT), is one way to infer causality. However, such experiments are not always possible or feasible.

To deal with issues of reverse causation and unobserved residual confounding, an approach commonly used in the economics and econometrics literature is that of Instrumental Variables (IV). This approach introduces a third variable (the instrument) that (partly) determines the level of the treatment or exposure of interest, but does not have a direct or indirect effect on the outcome variable, other than through its effect on treatment. This instrument can then be exploited to make causal inferences about the effect of the variable of interest on different outcomes.

'Mendelian randomization' refers to the random assignment of an individual's genotype at conception (Davey Smith and Ebrahim, 2003). Under certain assumptions that we discuss in detail below, observed associations between genetic variants and the outcome of interest cannot be due to reverse causation or confounding by behavioural or environmental factors, including those that occur *in utero*. Mendelian randomization can therefore be exploited to make causal inferences about the effects of modifiable (non-genetic) risk factors, on different outcomes (see Appendix A for a brief guide to the terms used in genetic studies).

Mendelian randomization is receiving increasing attention from epidemiologists, statisticians, economists and social scientists. Statisticians have highlighted some of the implicit statistical assumptions commonly made in Mendelian randomization studies (e.g. Didelez and Sheehan, 2007; Didelez, Meng, and Sheehan, 2010). Studies in genetic epidemiology on the other hand, 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., 2008). These conditions, however, have not been disseminated widely in the economics and social science literature, but the increasing availability of biomedical information in social science datasets makes understanding them crucial to the successful use of genotypes as instruments for modifiable risk factors. The main contribution of this paper therefore, is to discuss the conditions as defined in the epidemiology literature and relate them to the IV assumptions as used in statistics and economics. We do this using the potential outcomes framework, building on the work by Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996), which has been of great importance in linking the econometric IV literature to the potential outcomes framework. Note, however, that instrumental variables have also been defined using different approaches to causal inference, including causal Directed Acyclic Graphs (DAGs) and structural equation models (see Hernàn and Robins (2006) for an overview and comparison).

To clearly communicate best practice in genetic epidemiology to a wider social science audience, we review these conditions in the context of an empirical social science application. Specifically, we examine whether child adiposity (fat mass) causally affects academic achievement, using recently identified genetic variants as instrumental variables for adiposity.

We select this example for the following reasons. A causal effect of adiposity on academic achievement may run via various pathways. For example, obese children are more likely to be absent from school, to have sleep disorders, and be treated differently by teachers, parents and peers. All of these may affect children's (learning) environment and educational outcomes. However, an observed association between adiposity and academic achievement is not necessarily causal. It could, for example, be driven by children changing their eating habits *in response to* poor school results. In addition, one can never be sure that all relevant confounders are accounted for. In other words, the association may be driven by other unobserved factors relating to both weight and academic outcomes.

We attempt to deal with this possible reverse causation and unobserved confounding using Mendelian randomization, aiming to identify the causal effect of adiposity on academic achievement. When using carefully selected genetic variants as instrumental variables for adiposity, we find no evidence of a causal relationship between adiposity and academic performance, although the parameters are imprecisely estimated. In contrast, the use of ordinary least squares (OLS) leads to inferences of an inverse association.

The next section presents our statistical framework, details the conditions that need to be met for genetic variants to be used as instruments, presents our choice of genetic variants and contrasts this with the existing literature. Section three introduces the data. The results are presented in section four; section five concludes and discusses the implications of our findings.

2. The Use of Genetic Variants as Instrumental Variables

2.1 The Potential Outcomes Framework

Let S, A and Z denote random variables representing, respectively, the educational outcome, the adiposity measure and the (for now: binary) genetic variant as instrumental variable. $Z_i = 1$ indicates that individual i carries the genetic variant, $Z_i = 0$ implies that individual i does not carry the genetic variant.

Let $A_i(z)$ be the potential adiposity for individual i when the instrument is set to z. As with all potential outcomes, only one of the treatment assignments is ever observed for any one individual. Hence, we either observe i's adiposity when i does not carry the risk allele, $A_i(0)$, or we observe i's adiposity when i does carry the risk allele, $A_i(1)$. Similarly, let $S_i(a,z)$ be the potential outcome for individual i that would be obtained if the adiposity, the treatment variable, was set to a and the instrument set to z by external intervention. We refer to $A_i(z)$ and $S_i(a,z)$ as the potential treatments and potential outcomes respectively.

The individual treatment effect, or causal effect, is $S_i(a',z) - S_i(a,z)$, where a is some baseline value. Under the exclusion restriction discussed below, we can write $S_i(a',z) = S_i(a')$. Our causal estimand of interest can therefore be written as:

$$E[S_i(a') - S_i(a)]. \tag{1}$$

Hernàn and Robins (2006) outline the assumptions under which this causal effect is identified by the simple IV estimator that we use. These include linearity of the causal effect of A on S and no effect modification of the instrument on the treatment effect. Angrist, Graddy and Imbens (2000) however, specify the conditions under which the simple IV estimator identifies a weighted average of the derivative function of the (nonlinear) causal response function, exploiting a monotonicity assumption of the effect of the instrument on the exposure. We discuss these assumptions in turn.

Assumption 1. (Stable Unit Treatment Value Assumption, or SUTVA).

SUTVA implies that individual i's potential treatment is unrelated to the instrument values of other individuals, and that individual i's potential outcomes are unrelated to the treatment and instrument values of other individuals. In other words, there is no interference between units.

Assumption 2. (Independence).

$$Z_i \perp \{S_i(a,z), A_i(z)\}_{a,z}$$

The independence assumption implies that the instrument is independent of all potential outcomes and potential treatments, for all values of a and z. In other words, the instrument is as good as randomly assigned.

Assumption 3. (Exclusion).

$$S_i(a,1) = S_i(a,0), \text{ for all } a.$$

Exclusion implies that the potential outcomes, at any level of adiposity a, are unchanged by the presence or absence of the genetic variant. In other words, the only way through which the instrument affects the potential outcome is via A. Exclusion is distinct from the independence assumption, in that it is a claim about a unique channel of the causal effect of the instrument. It implies that $S_i(a,z)$ is a function of a only, and hence we can write $S_i(a,z) = S_i(a)$, as used in (1).

Assumption 4. (Non-zero effect of instrument on treatment).

$$E[A_i(1) - A_i(0)] \neq 0$$

This implies that expected potential adiposity is affected by the genetic variant and therefore that the instrument has an effect on the treatment.

Assumption 5. (Monotonicity).

$$P[A_i(1) \geq A_i(0)] = 1, \text{ for all } i$$

(or vice versa), saying that the potential treatment (i.e. adiposity) for individual i with the genetic variant is at least as high as the potential treatment for the same individual without the genetic variant for all i. In other words, all those affected by the instrument are affected in the same way.

In the presence of heterogeneous responses, the potential outcome for individual i can be written as a general function of a, say $S_i(a) \equiv g_i(a)$. We use assumptions 1 to 5 to interpret differences in average outcomes and treatments at different values of the instrument. Under these assumptions, the instrumental variables estimand, defined as the ratio of the difference in average outcomes at two values of the instrument to the difference in average treatment at the same two values of the instrument, can be written as:

$$\frac{E[S_i|Z_i = 1] - E[S_i|Z_i = 0]}{E[A_i|Z_i = 1] - E[A_i|Z_i = 0]}$$

$$= \frac{\int E\left[g_i'(q) | A_i(0) < q < A_i(1)\right] P\{A_i(0) < q < A_i(1)\} dq}{\int P\{A_i(0) < q < A_i(1)\} dq}, \qquad (2)$$

where $g_i'(q)$ is the derivative of $g_i(a)$ w.r.t. a evaluated at q. Hence, the IV estimand is a weighted average of the derivative function (Angrist, Graddy and Imbens, 2000; Angrist and Pischke, 2009).

In our application, the IV (Wald) estimand in (2) can be interpreted as the effect of adiposity on educational attainment for those whose adiposity is affected by the genetic instrument. This specific group is also referred to as the *compliers*: those who are induced to take treatment by assignment to treatment. In addition to these *compliers*, three further (latent) groups have been defined: those who do the opposite of their assignment (*defiers*), those who never take treatment, whatever their assignment (*never-takers*), and those who always take treatment, regardless of their assignment (*always-takers*). This stratification into latent groups is commonly used in the econometrics literature (see e.g. Angrist, Imbens and Rubin, 1996). Frangakis and Rubin (2002) have generalised this concept to any type of post-treatment variable within the potential outcomes framework, known as 'principal stratification'. Principal stratification has two important properties: (1) the strata (in this case: compliers, defiers, always-takers and never-takers) are not affected by the assigned treatment, and (2) comparisons of potential outcomes under different assignments *within* principal strata, also called principal effects, are well-defined causal effects. Hence, principal stratification can be seen as an extension of the IV methodology used here (see e.g. Barnard et al. (2003) for an application of principal stratification to a 'broken' randomized experiment, and Shinohara et al. (2009) and Shinohara and Frangakis (2009) for a comparison of principal stratification and IV in case-control studies).

The above discussion uses a binary instrument throughout. In our application however, we have multiple and multi-valued instruments. In the case of such discrete instruments, the number of instruments is irrelevant; it is the number of distinct values of the instrument vector that matters, averaging the (pairwise) instrument-specific weighted averages of the derivative function (Angrist, Graddy and Imbens, 2000).

2.2 Mendelian Randomization

We discuss the use of Mendelian randomization from a statistics and economics perspective in the context of a social study, with the aim of making causal inferences of the effect of a modifiable risk factor on an outcome of interest. As opposed to observational studies that estimate the effect of *short term* (current) differences in exposure, estimates from Mendelian randomization experiments reflect *lifetime* differences in exposure. Hence, they indicate the long term (cumulative) effects of the modifiable risk factor on the outcome of interest (Davey Smith and Ebrahim, 2005).

The concept of Mendelian randomization is closely linked to Randomised Controlled Trials (RCTs), where the allocation of treatment is randomised over all eligible individuals. Indeed, the instrumental variable methodology can be applied to analyse encouragement designs (such as RCTs where the treatment is the encouragement to participate) that are affected by non-ignorable non-compliance. Non-compliance refers to the fact that individuals can choose treatments other than those to which they are randomised. Non-ignorable non-compliance refers to participants choosing treatment 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 – adiposity – through lifestyle choices, such as diet and physical activity. Even after adjusting for baseline characteristics, these choices are likely to be related to their study outcome (educational attainment). In other words, they are non-ignorable. Mendelian randomization exploits the fact that there is an equal probability that either parental allele is transmitted to offspring. Whilst this allocation is random at the family trio level, at a population level it has been demonstrated that genetic variants are largely unrelated to the many socioeconomic and behavioural characteristics that are closely linked with each other and that confound conventional observational studies (Bhatti et al., 2005; Davey Smith et al., 2008; Kivimäki et al., 2008; Lawlor et al., 2008; see also Fisher, 1952 and Box 2010). Hence, as genes are randomly assigned during meiosis, individuals of different genotypes are expected not to differ systematically in any other respect.

We use Mendelian randomization and IV in the context of an empirical social science application. Although IV methods are widely used in statistics and economics, the use of genetic variants in these fields is relatively new. As noted by Didelez and Sheehan (2007), the potential limitations of Mendelian randomization experiments can be represented in two main categories. First, limitations related to the implicit statistical assumptions common in many Mendelian randomization studies (see e.g. Didelez and Sheehan (2007) and Didelez, Meng, and Sheehan (2010) for a thorough discussion). Second, potential limitations relating to the assumptions of the validity of the instrument. Genetic epidemiology studies mainly emphasize the latter: the importance of carefully examining several situations and (biological) processes that may violate the assumptions mentioned in section 2.1 (see e.g. Davey Smith and Ebrahim, 2003; Sheehan, et al. 2008; Lawlor et al., 2008). The existing research in economics however, has mainly failed to recognise these. But the increasing availability of biomedical information in social science datasets makes understanding of these conditions crucial to the successful use of genotypes as instruments for modifiable risk factors. We therefore discuss the concepts defined in epidemiology and relate them to the assumptions that need to be met to obtain causal estimates of the effect of the risk factor on the outcome of interest, as defined in the statistics and economics literature.

2.3 Conditions for the Use of Genetic Variants as Instrumental Variables

For a valid causal interpretation of the IV (Wald) estimand, we require assumptions 1 to 5 in section 2.1 to hold. Although we believe these assumptions are plausible, there are various situations that may violate them, and therefore need to be examined.

Assumption 1 – SUTVA

The potential treatment of any individual is not affected by the genetic variants of other individuals. Similarly, the potential outcome of any individual is not affected by the genetic variant or treatment of other individuals. The assumption of no interference between units is conventional in statistics and medical research, but the existence of peer effects may invalidate this. However, as peer effects are more likely to be due to (unobserved) behaviour than genes, this is less likely to be a problem here.

Assumption 2 – Independence

Although genotypes are randomly allocated at conception, the allele distribution may differ for different population subgroups. If these subgroups also have systematically different outcomes of interest, this could lead to an association between the two at the population level without an actual causal relationship. A systematic relationship between the allele frequency and the outcome of interest across different sub-populations is also referred to as population stratification. For example, allele frequencies can vary across ethnic groups. Any systematic differences in educational outcomes across these subpopulations that are not due to a genetic make-up may therefore lead to biased estimates of the effect of adiposity by violating the independence Assumption 2. In other words, despite the fact that genotypes are randomly allocated and with that satisfy the independence assumption, any population stratification can violate this assumption. This can be dealt with however, by examining the question of interest *within* ethnic groups, separately analyzing the different sub-populations, and/or adjusting for principal components from genome wide data that function as ancestry markers, relying on the conditional independence assumption.

As genotypes are randomly assigned *given* the parental genes, the presence of assortative mating based on genes can violate the independence assumption. The cleanest experiment, therefore, is one with biological siblings, examining randomization of genes within families (Davey Smith and Ebrahim, 2003). However, even when we only observe one individual per family, Mendelian randomization is valid assuming that, at the population level, genetic variants are unrelated to other characteristics, as has been shown in many studies (Bhatti et al., 2005; Davey Smith et al., 2008; Kivimäki et al., 2008; Lawlor et al., 2008; see also Fisher, 1952; Box 2011). One way to indirectly test independence is by exploring whether the distribution of individual or family-level characteristics that are available in the data is the same in different groups defined by the value of the instrument (i.e. different genotypes). The idea is that, if the instrumental variable is indeed randomized, there should be no systematic variation in the covariates by genotype, whether we use a within- or between-family analysis. This raises the question however, about *which* covariates to test for, as any characteristic is, in principle, a post-treatment variable with respect to the instrument. Hence, any systematic variation in these indirect tests does not necessarily indicate a violation of independence (or exclusion). It may be, for example, that the instrument is picking up additional causal effects of the same risk factor, or that it is picking up reverse causation from the outcome to a different covariate.

Assumption 3 – Exclusion

There can, in principle, be various situations in which Assumption 3 – the exclusion restriction – does not hold, invalidating the instruments. A first situation is that 'behaviours' may be affected by the genotype. As individuals inherit their genes from their parents, it may be important to consider whether parents' behaviours or preferences are affected by their genotype (and hence their offspring's genotype). This may be a problem in studies that examine maternal behaviours that influence the outcome of interest via intrauterine effects, or in studies where parental behaviours influence the outcome of interest via affecting their (child's) behaviour. For example, if one were interested in the effect of an individual's alcohol consumption on their later liver disease using a genetic variant that robustly relates to alcohol intake, any intrauterine effects of maternal alcohol consumption during pregnancy on offspring liver development could result in a violation of Assumption 3. This is because the mother's genotype is related to the offspring's genotype (the instrument), and will influence her alcohol consumption throughout life, including when she was pregnant. If maternal prenatal alcohol intake affects liver development *in utero* and its functioning later in the child's life, there is a link from the offspring's genotype (IV) to the outcome via an intrauterine mechanism. Similarly, if one is interested in the effect of adiposity (fat) on education, as we are below, parents who carry 'fat' alleles may be discriminated against in the labour market because of their higher weights. If this affects their behaviour or preferences for her child's education, Assumption 3 may be violated. The extent of this potential violation will depend on the effect sizes of the variants. In other words, if the variants shift the fat mass distribution by a modest amount, it may be unlikely to lead to strong (parental) responses.

A second situation relates to the mechanisms through which genetic variants affect the modifiable risk factor. These are often unknown. If the mechanism involves changes in behaviour or preferences that in addition to affecting the treatment also directly affect the outcome, Assumption 3 will be violated. If the mechanism only results in changes to the risk factor (i.e. they do not directly affect the outcome), Assumption 3 will not be violated.

Thirdly, the genetic instrument may be related to other genetic variants that affect the outcome of interest. Mendel's second law states that the inheritance of one trait is independent of the inheritance of another. However, it has been shown that this does not always hold and that some variants are likely to be co-inherited. This so-called 'Linkage Disequilibrium' (LD) does not occur for genetic variants on different (non-homologous) chromosomes, and the degree of LD is partly a function of the distance between the loci (see Appendix A for some of the genetic terms used here). Depending on the effects of the co-inherited variant, LD can bias the estimates. If our instrument Z is in LD with another polymorphic locus that affects only the modifiable risk factor A, the IV estimates remain consistent. However, if Z is in LD with a polymorphic locus that directly affects the outcome S, Assumption 3 is violated and the estimate will be biased and inconsistent. Relatedly, there is the situation of pleiotropy, where one genetic variant has multiple functions. The case is similar to that of LD, and will invalidate the IV approach if the pleiotropic effect influences the outcomes S directly, but not if it affects only other characteristics that are unrelated to the outcome of interest.

Fourth, a biological process that may bias causal estimates in Mendelian randomization studies is canalisation. This refers to a situation by which potentially disruptive influences of the risk factor on the outcome are buffered by foetal or post-natal developmental processes. For example, if an individual has a genetic variant associated with higher blood pressure, their arteries may develop to become resistant to the adverse effects of high blood pressure. This is difficult to test for, as the genetic variant may still be related to blood pressure, but any adverse health outcomes normally caused by higher blood pressure would not occur. Hence, canalisation can alter the association between genotype and outcome, without any change in the genotype-risk factor relationship. In other words, canalisation results in an indirect 'effect' of the genotype on the outcome of interest, violating the exclusion restriction.

Assumption 4 – Non-zero effect of Z on A

Mendelian randomization can only be used with genetic variants that have been robustly shown to affect the risk factor of interest. This means it relies on *prior* knowledge about the association between genotype and phenotype, as shown in a large number of independent studies. This latter point is especially important, as many initial genetic associations fail to replicate (Colhoun, McKeigue and Davey Smith, 2003). Without a robust and consistent population association, even if a significant *sample* correlation exists, Assumption 4 may be violated. Any correlation may simply be due to factors such as biased measurement (in genotype or phenotype) or chance.

However, even if a suitable and robust genetic instrument is available, it may explain little of the variation in observed phenotype. A weak association could result in a biased IV estimate and has implications for statistical power. If the alleles shift the distribution of A by a very small amount, the effect of A on S is identified only by this small difference in mean treatment, emphasizing the need for very large sample sizes, especially when the average causal effect of the risk factor on the outcome could be small. This, of course, is not a problem specific to Mendelian randomization, but refers to a more general problem of weak instruments, often encountered in IV studies (e.g. Angrist and Krueger, 1991).

Assumption 5 – Monotonicity

With random allocation of genetic variants and the fact that individuals do not know their genotypes, we assume that an individual who carries the risk allele is at least as heavy as the same individual, had she not carried the risk allele, satisfying the monotonicity Assumption 5. As this relies on knowing each individual's counterfactual adiposity, this remains an assumption. The literature only tells us that, at a group or population level, those who possess the genetic variant are heavier than those who do not. The monotonicity assumption could for example be violated in the presence of gene-environment interactions (i.e. when the effect of the environment on weight differs depending on the individuals' genetic predisposition, or when individuals' genetic predispositions are expressed differently in different environments).

2.4 The Role of Covariates in Mendelian Randomization Experiments

There is a large literature on the use of covariates in IV. Social and economics applications of IV generally include a wide set of control variables; the main motivation being that the *conditional* independence and exclusion restriction are more likely to be valid. A second reason for including covariates in many social and economics applications of IV is that it may reduce the variability in the dependent variable, leading to more precise estimates. Most IV models that include covariates use parametric specifications, but a causal interpretation of the IV estimate usually requires the assumptions of additive separability and constant treatment effects to hold. There are ways to relax these assumptions though, using non-parametric specifications (see e.g. Fröhlich (2007) for a review).

The situation is somewhat different, however, in RCTs/encouragement designs, and Mendelian randomization studies. When covariates enter the assignment mechanism in RCTs/encouragement designs, such as when randomization takes place within certain strata, these covariates should be controlled for, relying on the *conditional* independence assumption. The inclusion of further baseline covariates may in addition increase the precision of the estimates. In Mendelian randomization studies however, there are no baseline covariates. In addition, as covariates do not enter the assignment mechanism, assignment is independent of covariates, and we can rely on the (unconditional) independence assumption. Only in the presence of population stratification should the analyses be done *within* population subgroups, or should we be adjusting for principal components from genome wide data that function as ancestry markers, i.e. relying on the conditional independence assumption. Therefore, conditioning on covariates in not necessary in Mendelian randomization experiments.

Although one may choose to adjust for covariates to increase precision, this raises the issue of *which* covariates to include in a Mendelian randomization study, as any characteristic is, in principle, a post-treatment variable with respect to the instrument, and – with that – may be affected by the instrument. If the instrumented risk factor has multiple causal effects, or if the outcome of interest has a causal effect of its own, adjusting for such post-treatment variables may lead to biased estimates of the causal effect of interest. Indeed, we should not control for any 'downstream' behavioural covariates that are potentially affected by the treatment or outcome. Hence, although our application is in the context of a social study, we do not control for covariates, though we report the estimates that adjust for covariates in the text. Under the independence assumption, and in a situation where the instrumented risk factor and outcome do not (directly or indirectly) affect these covariates, the unadjusted and adjusted IV estimates should be similar, though the latter may be more precise.

2.5 Testing the Exclusion Restriction

There is no *direct* test for the validity of the exclusion restriction (Assumption 3). In other words, its validity will never be known with complete certainty and can only be examined indirectly or falsified by the data. To this end, however, Mendelian Randomization is no different from any other (non-genetic) IV study; the exclusion restriction *always* remains an assumption. Similar to other IV studies that use multiple instruments however, Mendelian randomization allows for potential violations through pleiotropy or LD (though not necessarily canalisation) to be tested when data is available on a large number of genetic instruments. More specifically, if multiple IV models - each using different combinations of these variants - predict the same causal effect, this is very unlikely to be due to some common pleiotropy or LD across the different sets of variants, assuming that the different variants are located on different chromosomes and affect the trait via different pathways (Davey Smith, 2011; Palmer et al., 2011). However, obtaining different causal effects with different combinations of variants does not *necessarily* point to a violation of the exclusion restriction, as variability in treatment effects may occur due to different compliant subpopulations for the different instrument sets (also known as different Local Average Treatment Effects, or LATEs).

In a constant effects model, one can indirectly test whether the exclusion restriction holds using an 'over-identification' test, provided there are more instruments than 'endogenous' variables (i.e. variables confounded by unobserved factors). Note, however, that this is not a test that the instruments are indeed valid. A problem with the over-identification test is that it has low power, especially when the underlying IV estimates are imprecise. In our framework of heterogeneous treatment effects however, over-identification tests are inappropriate, even when the underlying estimates are precise, as a rejection of the test need not imply a violation of the exclusion restriction. As discussed above, it may point to treatment effect heterogeneity, as different valid instruments for the same causal effect may estimate different parameters, with the final IV estimate being a weighted average of the different treatment effects (Angrist, Graddy and Imbens, 2000; Angrist and Pischke, 2009). Hence, although we report the test below, we cannot necessarily interpret it in a heterogeneous treatment effects framework.

Note that, although canalisation refers to a violation of the exclusion restriction, it cannot necessarily be tested (either directly or indirectly) using over-identification tests. Similar to the above, a rejection of the test cannot distinguish between treatment effect heterogeneity and canalisation. In fact, there is no (clear) way of testing or correcting for canalisation. However, for the complex traits that are largely of interest as causal risk factors in Mendelian randomization studies, there is no evidence that canalisation occurs in humans (Davey Smith and Ebrahim, 2003).

2.6 The Genetic Variants in our Empirical Application

We use two SNPs that have been consistently shown to relate to Body Mass Index (BMI, defined as weight in kg divided by height in metres squared) and adiposity in children and adults. Using a total of 38,759 individuals aged between 7 and 80 from 13 different cohorts of European ancestry, Frayling et al. (2007) explore the association between *FTO* and BMI, fat mass, the risk of being overweight and the risk of being obese. They find a positive association between the risk allele (A) and all measures of adiposity for individuals in all cohorts, in all countries, of all ages and of both sexes, with no difference between males and females. They show that *FTO* is specifically associated with fat mass, with weaker associations with lean mass. In addition, there is no association between *FTO* and birth weight, or *FTO* and height, suggesting that the relationship with BMI is largely driven by individuals' adiposity. They find that each copy of the risk allele is associated with an average increase in BMI of 0.2 units at age 7, to 0.4 units at age 11. For the average age-specific height, this refers to a weight increase of 0.3 and 0.9 kg respectively. As the genetic model for *FTO* is additive, meaning that each risk allele affects the phenotype by a similar amount, 11-year-olds who are homozygous for the rare allele (AA) are on average 1.8 kg heavier compared to those carrying no rare alleles. Using age-specific growth charts of weight, this corresponds to an increase from the median to the 58th percentile. However, there is much variation around this mean effect; the R^2 of a (linear) regression of adiposity on *FTO* is less than 1%.

Several different SNPs near *MC4R* have been associated with adiposity. We use the SNP identified by Loos et al. (2008). In addition to replicating the *FTO* findings, they find a positive relationship between rare allele (C) of *MC4R* and adiposity in genome-wide association data from 16,876 individuals and confirm this relationship in an additional 60,352 adults and 5,988 children. They find no differences by gender, and no effects on birth weight or children's height, again suggesting the association is mediated largely through an effect on adiposity. The genetic model for *MC4R* is dominant, meaning that the presence of any risk allele – either one or two – is associated with a similar increase in adiposity (Timpson et al., 2009). The findings on both *FTO* and *MC4R* have since been replicated in a vast number of other studies (see e.g. Hinney et al., 2007; Hunt et al., 2008; Willer et al., 2009; Thorleifsson et al., 2009; Meyre et al., 2009; Speliotes et al., 2010).

Our specific choice of genetic variants can be related to the assumptions for suitable use of genetic variants as instruments discussed in Section 2.1 and 2.3. First, population stratification due to ethnicity may violate the independence assumption, as *FTO*-allele frequencies are known to vary by ethnic group (Frayling et al., 2007). However, it is not likely to be a problem here, as our cohort is recruited from a specific geographically defined region with a predominantly white population. In addition, our analysis only includes children whose mother describes herself and the child's father as white.

Second, note that those who carry the rare allele of *FTO* and/or *MC4R* do not all become obese; the variants increase the average body weight by a modest amount. Hence, it is unlikely to observe any strong (parental) responses to increased body weight, such as changing children's diets.

Third, we rely on the theory of random allocation of genetic variants (independence) and on the empirical evidence that shows that genetic variants are unlikely to be related to unmeasured confounders (exclusion). Pleiotropy or LD would bias the IV estimates if the variant affects the outcome directly or if the linkage is with another variant that directly affects educational attainment.

Fourth, the possible mechanisms through which *FTO* and *MC4R* affect adiposity are increasingly studied in the medical literature. Although this work is ongoing, there is substantial evidence that the variants are associated with an increased consumption of fat and energy (see e.g. Timpson et al., 2008; Cecil et al., 2008). The literature suggests that the SNPs increase food intake due to diminished satiety (Wardle et al., 2008), rather than through pathways that affect our schooling outcome of interest, suggesting that Assumption 3 is satisfied.

Fifth, as noted above, the biological mechanisms through which *FTO* and *MC4R* affect adiposity are unknown, but the evidence to date suggests that size at birth is not affected by these variants. Therefore, canalisation during the foetal period is unlikely to be a problem.

Sixth, the prior findings of robust associations between the genetic variants and individual adiposity justify their use as instruments (Assumption 4). Each *FTO* risk allele leads to an average increase of 0.9 kg; carrying one or two *MC4R* risk alleles is related to an average increase of 0.6 kg. As mentioned above however, with much variation around this mean effect, our two genetic variants explain little of the total variation in adiposity: $R^2 < 1\%$. Using the standard statistical tests, we will examine the strength of our instruments in the application below.

Finally, as monotonicity (Assumption 5) can be violated by gene-environment interactions, we examine this indirectly, testing whether the association between adiposity, *FTO* and *MC4R* differs by 'environment', as defined by gender, social class, mother's education, and income. The results (available from the authors on request) show no significant differences, suggesting that these interactions do not play an important role for the genetic variants used here.

2.7 Main Advantages Relative to Previous Studies that use Genetic Variants

The existing economics literature includes three studies that exploit genetic variation to identify the effects of BMI on economic outcomes. Ding et al. (2009) examine the effects of several health conditions, one of which is BMI, on adolescent's academic achievement. Their IV results show large and significant negative effects on female's Grade Point Average (GPA), but not for males. GPAs for obese girls are on average 0.8 points lower than those for non-obese girls. They use four genetic variants as instrumental variables: the dopamine transporter (*DAT1*), the dopamine D2 receptor (*DRD2*), tryptophan hydroxylase (*TPH*) and cytochrome P4502B6 (*CYP2B6*). Fletcher and Lehrer (2011) take a similar approach to Ding et al. (2009), but use a different dataset (the Add Health data) to exploit within-family genetic inheritance. They find no evidence that obesity affects academic achievement. In addition to *DAT1* and *DRD2*, their instruments include the dopamine D4 receptor (*DRD4*), the serotonin transporter (*5HTT*), monoamine oxidase (*MAOA*) and cytochrome P4502A6 (*CYP2A6*). Finally, Norton and Han (2008) examine the effects of BMI on labour market outcomes using *DAT1* and *DRD4* as instrumental variables for BMI and find no evidence of a causal association.

The discussion in Section 2.3 above highlights the importance of the choice of genetic variants in Mendelian randomization experiments. Although the validity of the exclusion restriction can never be tested directly, it is very *un*likely that genes related to neurotransmitters, such as those used in the three studies described above, are valid instruments. The inherent problem is that neurotransmitters are implicated in many different neurological processes. Hence, it is difficult to argue that they can be used as valid instruments for one specific risk factor without being associated with others that could plausibly influence the outcome of interest (Cawley et al., 2011; von Hinke Kessler Scholder et al., 2011).

In addition, Assumption 4 states that consistent and robust associations should have been shown between the genotype and phenotype in a large number of independent studies. The three studies cited above do not appear to have taken this approach (Lawlor, Windmeijer and Davey Smith, 2008). Rather than basing their selection of genetic variants on associations that are robustly shown in the literature, their choice of instruments seems to be data-derived: using either forward stepwise estimation (Ding et al., 2009) or selecting those SNPs that have nominally statistically significant *sample* correlations in the first stage (Fletcher and Lehrer, 2011). In fact, both Ding et al. (2009) and Fletcher and Lehrer (2011) acknowledge that there is weak and inconsistent evidence in the medical literature, based on very small unrepresentative clinical samples, of the association between their genetic variants and health status or behaviours. Indeed, the IV strategy is invalid when relying only on these *sample* associations, as the absence of a robust population association violates Assumption 4. Norton and Han (2008) base their selection of SNPs on a study by Guo et al. (2006), who argue that there is a negative association between the D4.7/D4.7 genotype of *DRD4* and obesity. This relationship, however, has not been replicated in other independent studies (see for example

Hinney et al. (1999), or Fletcher and Lehrer (2011) who find an insignificant but *positive* association).

In addition, these studies are unable to replicate various associations they note are reported in the literature. For example, Ding et al. (2009) find no association between the number of 10-repeat alleles of the DAT1 gene and obesity, whilst they note the literature reports a positive relationship, and Norton and Han (2008) find a negative correlation. Fletcher and Lehrer (2011) fail to show any correlation between the A1A1 variant of *DRD2* and obesity. But given that the evidence of a robust association for these variants is lacking, this is not surprising (Lawlor, Windmeijer and Davey Smith, 2008). Furthermore, Norton and Han (2008) argue that the effects of the genetic variants differ by gender, while Patsopoulos et al. (2007) note that most claims of gender differences are spurious. Finally, Norton and Han (2008) use several variants as additional controls rather than instruments, as they fail the over-identification tests (*SLC6A4*, *MAOA*, *DRD2* and *CYP2A6*). Fletcher and Lehrer (2011) and Ding et al. (2009) use several of these as their instruments.

3. Data

Our data are from a cohort of children born in one geographic area (Avon) of England. Women eligible for enrolment in the population-based Avon Longitudinal Study of Parents and Children (ALSPAC) had an expected delivery date between 1 April 1991 and 31 December 1992. Approximately 85% of these mothers enrolled, leading to about 14,000 pregnancies. The Avon area has approximately 1 million inhabitants and is broadly representative of the UK as a whole, though slightly more affluent than the general population (Golding et al. 2001; see www.bristol.ac.uk/alspac for more a detailed description of the representativeness of the sample, its enrolment, and response rates). Detailed information on the study children and their families has been collected using a variety of sources, including self-completed questionnaires, data extraction from medical and educational records, in-depth interviews, and biological samples. Note however, that ALSPAC is a cohort; there is no systematic data collection on siblings that we can exploit.

A total of 12,620 children survived past the age of 1 and returned at least one questionnaire. Of these, 642 were excluded because either their mother or father is of non-white ethnic origin, leaving 11,978 potential participants. Our sample selection process is as follows. First, we select those children for whom we observe their genotypes, leaving us with 7,368 children. Second, we drop children with missing data on fat mass. Children were invited to attend research clinics, where their anthropometric measures were recorded. As not all children attended these clinics, our sample sizes reduce to just over 4,500. We further restrict the sample to those children for whom we observe their educational outcomes, leading to a final sample size of 3,729 children. *T*-tests of mean equality show the final sample to be slightly wealthier than the original ALSPAC sample, with mothers being somewhat older and having fewer mental health problems. The probability of being in the final sample however, is unrelated to *FTO* and *MC4R*, suggesting that sample selection is unrelated to the genotypes used here.

3.1 Measures of Academic Achievement

Our main outcome measure is the child's Key Stage 3 (KS3) score. The KS3 exam is a nationally set exam, taken by all 14-year-olds in English state schools. This measure of children's performance is therefore objective and comparable across all children. Children's scores for three subjects (English, maths and science) are obtained from the National Pupil Database, a census of all pupils in England within the state school system, which is matched into ALSPAC. We use an average score for the three subjects, standardised on the full sample of children for whom data is available, with mean 100 and standard deviation 10.

3.2 Measures of Child Adiposity and the Genetic Variants

Our main measure for child adiposity is the child's body fat mass (adjusted for age in months, height and height squared), as measured by a dual-energy X-ray absorptiometry scan (DXA) at age 11. This method scans the whole body, dividing it into body fat, lean tissue mass, and bone density. We standardise fat mass on the full sample of children for whom data are available, with mean 100 and standard deviation 10. For the genetic variants, we use two SNPs that have been consistently found to relate to weight: *FTO* (rs9939609) and *MC4R* (rs17782313; the rs-number is an identification tag that uniquely positions the polymorphism in the genome). Due to the nature of the association between *MC4R* and adiposity (a dominant genetic model), we group individuals who carry one or more risk alleles (C) together. Hence, we observe children in two groups based on their genotype: TT vs. CT/CC. *FTO* is specified as having three categories: no risk alleles (homozygous TT), one risk allele (heterozygous AT) and two risk alleles (homozygous AA).

3.3 Descriptive Statistics

As discussed in section 2.1, we observe six mutually exclusive instrumental variables, as defined by the combination of the number of rare alleles in *FTO* and *MC4R*. We begin by showing their frequencies. Without loss of generality, we order the instruments by their mean fat mass, as presented in Table 1. Hence, our instrument contains the following combinations of the genetic variants: $(FTO, MC4R) \in \{(TT, TT), (TT, CT/CC), (TA, TT), (TA, CT/CC), (AA, TT), (AA, CT/CC)\}$. We refer to these as $Z = 1, \dots, 6$ respectively. Table 1 shows that the first group defined by the instrument ($Z = 1$, i.e. those homozygous for the common allele of both *FTO* and *MC4R)* has a mean (standardised) adiposity of 98.4, with the last group ($Z = 6$) having a mean (standardised) adiposity of 102.4.

We can further examine the adiposity distribution for individuals within each of the six groups by plotting the empirical fat mass distribution functions (as in Angrist, Graddy and Imbens, 2000). Figure 1 shows large differences, with the adiposity distribution of the $(AA, CT/CC)$ group (i.e. $Z = 6$; those homozygous for the rare allele of *FTO,* and heterozygous or homozygous for the rare allele of *MC4R*) lying to the right of all others for most values of adiposity.

Figure 2 presents the differences between these distribution functions, plotting the unnormalised weight functions. As discussed in section 2.1, the IV approach estimates the coefficient of interest (2) as the average causal derivative for the shift in adiposity at each value of the instrument. We therefore show the weight functions for those combinations of Z used in the final IV estimate.

The figures show slightly different weight functions for different instrument values. In the left figure for example, taking the difference between the distribution function of $Z = 1$ and $Z = 2$, more weight is given to the higher end of the fat mass distribution, whereas the difference between $Z = 2$ and $Z = 3$ gives more weight to the lower end of the adiposity distribution. The other weight functions, shown in the figure on the right, show similar weights for the different instrument values, mainly affecting the middle of the adiposity distribution. The smallest weights in the IV estimation are given for the difference between the distribution functions of $Z = 4$ vs. $Z = 5$, the largest weights are for $Z = 5$ vs. $Z = 6$.

Finally, Appendix B indirectly tests the independence assumption by exploring whether the distribution of covariates is the same in different groups defined by the value of the instrument. Although there are no actual pre-treatment variables with respect to the instrument, and significant differences do not necessarily indicate violation of independence (see also section 2.3), we find no evidence of systematic differences for the different covariates, providing at least suggestive evidence of randomization of the genetic variants.

4. Results

Table 2 presents the results. Column 1 shows the association between the KS3 score at age 14 and fat mass at age 11. This estimate is similar to the adiposity-coefficient of an OLS regression of KS3 on fat mass. The (unadjusted) relationship between fat mass and educational attainment is negative, with a one standard deviation increase in fat mass associated with a 0.1 standard deviation decrease in test scores.

Column 2 presents the first-stage regression results, showing a strong positive relationship between the different instrument values and child fat mass. The estimates increase with the instrument values, as expected by construction of the instrument. The strength of the relationship between the instruments and fat mass is shown by the first stage *F*-statistic; with a value of 9.4, it is relatively weak. However, if we were to use Z as one instrument, rather than five separate dummies, the *F*-statistic rises to over 45, confirming the strength of our genetic variants.

The second stage IV results are presented in column 3, showing no effects of fat mass on educational performance. Although the IV estimate is of similar magnitude but opposite sign to that in column 1, the large standard errors preclude us from rejecting the null of no effect. However, with a *p*-value of 0.053 for the Hausman test, there is some support for the IV estimate as opposed to the OLS estimate, though any such judgement should be based on a synthesis of all the evidence, rather than on this one test alone. Appendix C presents the IV estimates that distinguish between the different values of the instrument Z that are shown in Table 1. As discussed in section 2.1, the final IV estimate in Table 2, column 3, is a weighted average of these separate regressions. The over-identification (Hansen J) test does not reject the null although, as discussed above, it is difficult to interpret this test in a heterogeneous treatment effects framework. Even when we control for the set of background characteristics mentioned in Appendix B, the IV estimate remains very similar to the unadjusted results (0.111 vs. 0.140 in Table 2), with a slightly smaller standard error (0.111 vs. 0.131), though the *p*-value remains large (0.315 vs. 0.282).

5. Conclusion and Discussion

The increasing availability of biomedical data, in combination with a growing medical literature on the effects of carrying specific genetic variants, introduces a different approach to the examination of certain risk factors on different outcomes. This paper discusses the method of instrumental variables using Mendelian randomization, and relates this to the statistical framework of potential outcomes.

We discuss the specific conditions that need to be met for genetic variants to be used as instruments, and relate these to the statistical assumptions necessary for identification of the average causal response using instrumental variables. These conditions have not been well defined in the current social science literature, but the increasing availability of biomedical data makes understanding of these conditions crucial to the successful use of genotypes as instruments for modifiable risk factors. To clearly communicate best practice in genetic epidemiology to a wider social science audience, we review these conditions in the context of an empirical social science application. Specifically, we examine whether child adiposity causally affects academic achievement, using recently identified genetic variants as instrumental variables for adiposity.

We argue that these variants are the best current candidates for use as genetic markers: they have been shown to be associated with adiposity in large population samples and we argue that they are likely to meet the conditions required for suitable instruments. We also use direct measures of body fat mass, rather than the generally used BMI. OLS shows that leaner children perform better in school tests compared to their more adipose counterparts. Our genetic IV analysis, however, shows no evidence that children's fat mass affects their academic performance, although the estimates are imprecise. Based on our robust IV approach though, we conclude that adiposity is not a major determinant of educational outcomes.

Our discussion of the conditions for the suitability of genetic variants as instruments and our application raise a more general issue of the use of genetic variants as instrumental variables. In our case, while our instruments are not overly weak in a statistical sense, their effects may be too small to impact on the possible pathways to academic performance. In other words, a two kilogram increase in adiposity may not lead to a large drop in self-esteem or an increase in absenteeism. To that end, it is perhaps not surprising that we do not find a significant effect on academic performance. That said, *FTO* is the strongest adiposity-marker yet identified. It is relatively unlikely that common variants will be found with larger adiposity-effects, as those with larger effects tend to be discovered before smaller ones (though rare variants with stronger effects may be identified).

This illustrates the more general question of whether genetic variants are powerful enough to identify causal effects in studies examining economics or social science outcomes. The answer to this question will depend on the variant, the risk factor and the outcome of interest. With a rapidly growing medical literature on the effects of carrying specific genetic variants, one option is to wait for more variants to be identified and to combine these into one or more instrumental variables, such as a (weighted) count of the number of risk alleles (see e.g. von Hinke Kessler Scholder et al., 2010). This could increase the explained phenotypic variation and with that, the precision of the estimates. But as noted, in the case of adiposity - and other physical attributes that economists have been interested in, such as height - any additional variants are likely to have even smaller effects than those already identified.

In conclusion, we argue that genetic instruments need to be used with care. Their appropriate use requires that several conditions, which have not hitherto been spelt out in the social science literature, are met. But even if these conditions are met, the sample sizes in data sets that contain both genetic markers and outcomes of interest to social scientists may be too small to obtain definitive results. Indeed, even with almost 4000 observations, our standard errors are relatively large. With a rapid increase in the number of genome wide association studies being done, and with a decrease in their costs however, this may change. With that, we believe that Mendelian randomization presents a promising approach to estimate the causal effect of a modifiable risk factor on one or more outcomes of interest, though we reiterate that this hinges on the correct selection of genotypes as instruments for risk factors.

Appendix A: A Brief Introduction to Genetics

Each cell in the human body contains a nucleus in which most DNA (99.9995%) is kept. DNA is stored in structures called chromosomes, where each chromosome contains a single continuous piece of DNA. All cells in the human body apart from gametes (i.e. germ cells) contain 46 chromosomes, organised into 23 chromosome pairs: one copy of chromosome 1-22 from each parent, plus an X-chromosome from the mother and either an X or a Y chromosome from the father.

Locations (or loci) where DNA varies between people are called polymorphisms. The most commonly studied form of polymorphism is a Single Nucleotide Polymorphism (SNP): a single base-pair variation in a DNA locus. As chromosomes come in pairs, humans have two base-pairs at each locus, called alleles. These alleles can either be the same or different. The term genotype is used to describe the specific set of alleles inherited at a particular chromosome locus. For example, individuals can have one of three genotypes of *FTO*: they can be homozygous for the common allele (TT), heterozygous (AT), and homozygous for the rare allele of *FTO* (AA). The visible or measurable effect of a particular genotype is called the phenotype.

The phenotype we examine is fat mass. Many studies have examined the heritability of adiposity, where heritability is defined as the proportion of the total variance that is explained by genetic factors; most commonly calculated from twin studies by comparing intra-pair correlations for a characteristic in monozygotic twins with that in dizygotic twins. These studies generally report large heritability estimates: between 0.4 and 0.7. A high heritability however, does not imply that any individual genetic variant has large phenotypic effects. For example, there are many different SNPs that affect human weight, though all with small effects: so-called ‘polygenes’. Together, these variants may have a large phenotypic effect.

Until recently, researchers mainly used a ‘candidate gene approach’ to examine associations between individual genetic variants and a phenotype. This approach consists of testing a specific hypothesis: based on biological knowledge, researchers examine the association between one particular variant (the candidate genetic variant) and a phenotype. These studies produced many false-positive findings (Colhoun et al., 2003) and were inefficient. Genome wide association studies (GWAS) followed, genotyping 500,000 to over 1,000,000 SNPs in one go and relating all SNPs to the phenotype of interest in a hypothesis-free way. Stringent criteria are used for GWAS *p*-values to take account of this hypothesis-free approach. Studies are either two-stage studies, where one or more GWAS is performed, after which the small number of SNPs that reach GWAS levels of statistical significance are typed in other independent samples to examine the robustness. Alternatively, studies consist of a number of independent GWAS containing a large total sample size, where only those SNPs that have consistent associations across all studies are interpreted as robust.

Appendix B: Indirect Test of Independence of Covariates and Genetic Variants

Table B1 presents the coefficients (standard errors) of a regression of the covariate presented in the first column on the five values of the instrument ($Z = 2,3,4,5,6$). The final column shows the *p*-value of an *F*-statistic testing whether the coefficients jointly equal zero. With random assignment of the genetic variants, there should be no systematic variation in the covariates by genotype. Although some *p*-values are lower than the commonly used 0.01 or 0.05, we find no evidence of systematic differences for the different covariates, providing at least suggestive evidence of randomization of genetic variants.

Table B1: An indirect test of independence: regressing the covariate on the instrument-dummies.

	$Z = 2$	$Z = 3$	$Z = 4$	$Z = 5$	$Z = 6$	*p*-value of F-test: instrument-coefficients jointly equal to zero
Girl	-0.043	-0.024	-0.010	0.062*	0.010	0.045
	(0.03)	(0.02)	(0.03)	(0.032)	(0.037)	
Birth weight (g)	-38.58	5.43	-17.79	8.76	-15.48	0.685
	(30.70)	(25.90)	(27.12)	(35.60)	(40.96)	
Age at KS3 (in months)	0.22	0.09	0.14	0.02	0.38	0.737
	(0.20)	(0.18)	(0.19)	(0.24)	(0.27)	
Ln(income)	-0.02	0.02	-0.01	0.02	0.01	0.716
	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)	
Mother’s education	0.05	0.04	0.02	0.05	-0.02	0.779
	(0.05)	(0.04)	(0.05)	(0.06)	(0.06)	
Raised by natural father	-0.01	0.01	-0.01	0.00	-0.00	0.600
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	
Social class at birth	-0.06	-0.06	-0.03	-0.18**	-0.19**	0.157
	(0.07)	(0.06)	(0.07)	(0.08)	(0.09)	
Mum works part-time	0.00	-0.01	0.02	0.01	-0.02	0.901
	(0.03)	(0.02)	(0.03)	(0.03)	(0.04)	
Mum works full-time	0.00	0.00	-0.02	0.02	-0.02	0.210
	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	
Partner employed	-0.02	0.01	-0.01	-0.01	-0.02	0.618
	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	
Index of Multiple	-0.93	0.59	0.06	0.26	0.55	0.437
Deprivation	(0.75)	(0.67)	(0.70)	(0.90)	(1.01)	
Alcohol during pregnancy	-0.02	0.04	0.01	-0.02	0.04	0.161
	(0.03)	(0.02)	(0.03)	(0.03)	(0.04)	
Smoke during pregnancy	0.01	0.01	0.01	-0.03	-0.01	0.615
	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	
Breastfeeding	0.10	0.06	0.12**	-0.03	0.22**	0.053
	(0.07)	(0.06)	(0.06)	(0.08)	(0.09)	
Mother’s age at birth	-0.01	0.03	0.01	-0.03	-0.13*	0.265
	(0.05)	(0.04)	(0.05)	(0.06)	(0.06)	

Notes: The *p*-value in the last column corresponds to an *F*-test of the coefficients on the five instrumental variables jointly equalling zero. * p<0.1; ** p<0.05; *** p<0.01; Ln(income) is measured when the child is aged 3-4 as is in 1995 prices; mother’s educational level is a categorical variable with four values (less than ordinary (O) level, O-level, advanced (A) level, and university degree); social class is measured using the standard UK classification of class based on occupation (professional (I), managerial and technical (II), non-manual skilled (IIInm), manual skilled (IIIm), semi-skilled (IV) and unskilled (V)); the Index of Multiple Deprivation (IMD) is a measure of local area deprivation, referring to areas containing about 8000 persons; breastfeeding is a categorical variable (never, <1 month, 1-3 months and 3+ months).

Appendix C: IV Estimation using Different Instrument Sets

Table C1 presents the IV results using different instrument sets. As discussed in section 2.1, we recode our multiple multi-valued instruments to a set of mutually exclusive dummy variables, resulting in six binary instrumental variables. Here we show the IV estimates using each of these distinct values.

Column (1) replicates the IV estimate from Table 2, using the five dummies simultaneously. Column 2 shows the IV estimate when using only those observations with the instrument Z equal to 1 or 2. Column 3 includes those with $Z = 2$ or 3; column 4 uses those with $Z = 3$ or 4; column 5 includes those with $Z = 4$ or 5; and column 6 uses those with $Z = 5$ or 6. Finally, column 7 shows the IV estimate using a linear specification of the instrument.

This shows that, for low values of the instrument $(Z = 1{,}2{,}3)$, the IV estimate is negative, though always with large confidence intervals. Higher values of Z give positive estimates, with a very large estimate for $Z \in \{4{,}5\}$ in column 5, but it is very imprecisely estimated. Note also that, as shown in Figure 2, this estimate gets the lowest weights over the distribution of fat mass, and hence the final IV estimate in Table 2 is not driven by this large estimate. In contrast, the largest weights shown in Figure 2 are for the specification with $Z \in \{5{,}6\}$ (column 6 below), which is indeed closest to our final IV estimate.

Table C1. IV estimates of the average response in standardised KS3 to a 1 standard deviation change in fat mass, different instrument sets

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Instruments:	Full sample, five instruments	$Z \in \{1{,}2\}$	$Z \in \{2{,}3\}$	$Z \in \{3{,}4\}$	$Z \in \{4{,}5\}$	$Z \in \{5{,}6\}$	Linear specification of Z
Fat Mass	0.140	-0.204	-0.107	0.131	1.876	0.139	0.136
	p=0.282	*p=0.801*	*p=0.821*	*p=0.836*	*p=0.660*	*p=0.785*	*p=0.305*
	[-0.12, 0.40]	[-1.79, 1.38]	[-1.03, 0.82]	[-1.11, 1.37]	[-6.48, 10.2]	[-0.86, 1.13]	[-0.12, 0.40]
N	3729	1356	1604	1784	1103	589	3729

Notes: Column 1 is the specification as in Table 2, column 3: using the five instruments $Z \in \{2{,}3{,}4{,}5{,}6\}$. Column 2 only uses (TT, TT) and (TA, TT) (i.e. Z equalling 1 or 2); Column 3 only uses observations with Z equalling 2 or 3. Column 4 includes those with $Z = 3$ or 4; Column 5 includes those with $Z = 4$ or 5; Column 6 includes those with $Z = 5$ or 6. Column 7 shows the IV estimate using a linear specification of the instrument. 95% confidence intervals in square brackets; *p* is p-value for standard t-ratio.

References

Angrist, J. D., Pischke, J-S. (2009) *Mostly Harmless Econometrics: An Empiricist's Companion,* Princeton: Princeton University Press.

Angrist, J. D., Graddy, K., Imbens, G. (2000) The Interpretation of Instrumental Variables Estimators in Simultaneous Equation Models with an Application to the Demand for Fish. *Review of Economic Studies,* **67**, 499-527.

Angrist, J. D., Imbens, G.W., Rubin, D.B. (1996) Identification of Causal Effects using Instrumental Variables (including discussion). *Journal of the American Statistical Association*, **91**(434), 444-472.

Angrist, J.D., Krueger, A. (1991) Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics,* **106**, 979-1014.

Barnard, J. et al. (2003) Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City. *Journal of the American Statistical Association*, **98**(462), 299-312.

Bhatti, P., Sigurdson, A.J., Wang, S.S., Chen, J., Rothman, N., et al. (2005) Genetic Variation and Willingness to Participate in Epidemiologic Research: Data from Three Studies. *Cancer Epidemiology Biomarkers and Prevention,* **14**, 2449-2453.

Box, J.F. (2010) Commentary: On RA Fisher's Bateson Lecture on Statistical Methods in Genetics. *International Journal of Epidemiology*, **39**, 335-339.

Cawley, J., Han, E., Norton, E. (2011) The Validity of Genes Related to Neurotransmitters as Instrumental Variables. *Health Economics*, **20**(8), 884-888.

Cecil, J., Tavendale, R., Watt, P., Hetherington, M.M., Palmer, C.N.A. (2008) An obesity-associated FTO gene variant and increased energy intake in children. *The New England Journal of Medicine,* **359,** 2558-2566.

Colhoun, H., McKeigue, P., Davey Smith, G. (2003) Problems of Reporting Genetic Associations with Complex Outcomes. *The Lancet,* **361**, 865-872.

Davey Smith, G., Ebrahim, S. (2003) 'Mendelian Randomization': Can Genetic Epidemiology Contribute to Understanding Environmental Determinants of Disease? *International Journal of Epidemiology,* **32**, 1-22.

Davey Smith, G., Ebrahim, S. (2004) Mendelian Randomization: Prospects, Potentials, and Limitations. *International Journal of Epidemiology*, 33, 30-42.

Davey Smith, G., Ebrahim, S. (2005) What can Mendelian Randomisation tell us about Modifiable Behavioural and Environmental Exposures? *BMJ* **330**, 1076-1079.

Davey Smith, G., Lawlor, D.A., Harbord, R., Timpson, N., Day, I., Ebrahim, S. (2008) Clustered Environments and Randomized Genes: A Fundamental Distinction between Conventional and Genetic Epidemiology. *PLoS Medicine,* **4**, 1985-1992.

Davey Smith, G. (2011) Use of Genetic Markers and Gene-Diet Interactions for Interrogating Population-Level Causal Influences of Diet on Health. *Genes and Nutrition,* **6**(1), 27-43.

Didelez, V., Sheehan, N.A. (2007) Mendelian Randomization as an Instrumental Variable Approach to Causal Inference. *Statistical Methods in Medical Research*, **16**, 309-330.

Didelez, V., Meng, S., Sheehan, M.A. (2010) Assumptions of IV Methods for Observational Epidemiology. *Statistical Science*, **25**(1), 22-40.

Ding, W., Lehrer, S.F., Rosenquist, N.J., Audrain-McGovern, J. (2009) The Impact of Poor Health on Academic Performance: New Evidence using Genetic Markers. *Journal of Health Economics,* **28**, 578-597.

Fisher, R.A. (1952) Statistical Methods in Genetics. *Heredity,* **6**, 1-12. Reprinted: *International Journal of Epidemiology*, *2010; 39:329-335*

Fletcher, J.M., Lehrer, S.F. (2011) Genetic Lotteries within Families. *Journal of Health Economics*, **30**, 647-659.

Frayling, T., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C., et al. (2007) A Common Variant in the FTO Gene is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science,* **316**, 889-894.

Frangakis, C.E., Rubin, D.B. (2002) Principal Stratification in Causal Inference. *Biometrics*, **58**, 21-29.

Fröhlich, M. (2007) Nonparametric IV Estimation of Local Average Treatment Effects with Covariates. *Journal of Econometrics*, **139**, 35-75.

Golding, J., Pembrey, M., Jones, R., ALSPAC Study Team. (2001) ALSPAC - The Avon Longitudinal Study of Parents and Children: I. Study Methodology. *Pediatric and Perinatal Epidemiology,* **15**, 74-87.

Guo, G., North, K., Choi, S. (2006) DRD4 Gene Variant associated with Body Mass: The National Longitudinal Study of Adolescent Health. *Human Mutation,* **27**, 236-241.

Hernàn, M.A., Robins, J.M. (2006) Instruments for Causal Inference: An Epidemiologist's Dream? *Epidemiology,* **17**, 722-729.

von Hinke Kessler Scholder, S., Davey Smith, G., Lawlor, D.A., Propper, C., Windmeijer, F. (2011) Mendelian Randomization: The Use of Genes in Instrumental Variable Analyses. *Health Economics*, **20**(8), 893-896.

von Hinke Kessler Scholder, S., Davey Smith, G., Lawlor, D.A., Propper, C., Windmeijer, F. (2010) Child Height, Health and Human Capital: Evidence using Genetic Markers. *CMPO Working Paper 10/245*.

Hinney, A., Schneider, J., Ziegler, A., Lehmkuhl, G., Poustka, F., Schmidt, M.H., et al. (1999) No Evidence for Involvement of Polymorphisms of the Dopamine D4 Receptor Gene in Anorexia Nervosa, Underweight, and Obesity. *American Journal of Medical Genetics,* **88**, 594-597.

Hinney, A., Nguyen, T.T., Scherag, A., Friedel, S., Bronner, G., Muller, T.D., et al. (2007) Genome Wide Association (GWA) Study for Early Onset Extreme Obesity Supports the Role of Fat Mass and Obesity Associated Gene (FTO) Variants. *PLoS ONE,* **2,** e1361.

Holland, P. (1986) Statistics and Causal Inference. *Journal of the American Statistical Association*, **81**, 945-970.

Hunt, S.C., Stone, S., Xin, Y., Scherer, S.A., Magness, C.L., Iadonato, S.P., Hopkins, P.N., Adams, T.D. (2008) Association of the FTO Gene with BMI. *Obesity,* **16**, 902-904.

Imbens, G.W. Angrist, J.D. (1994) Identification and Estimation of Local Average Treatment Effects. *Econometrica*, **62**(2). 467-475.

Kivimäki, M., Davey Smith, G., Timpson, N.J., Lawlor, D.A., Batty, G.D., et al. (2008) 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. *European Heart Journal,* **29**, 2552-2560.

Lawlor, D.A., Windmeijer, F., Davey Smith, G. (2008) Is Mendelian Randomization "Lost in Translation"? *Statistics in Medicine,* **27**, 2750-2755.

Lawlor, D., Harbord, R.M., Sterne, J.A., Timpson, N.J., Davey Smith, G., (2008) Mendelian Randomization: Using Genes as Instruments for Making Causal Inferences in Epidemiology. *Statistics in Medicine,* **27**, 1133-1163.

Loos, R.J., Lindgren, C.M., Li, S., Wheeler, E.,Zhao, J.H., Prokopenko, I., et al. (2008) Common Variants Near *MC4R* are Associated with Fat Mass, Weight and Risk of Obesity. *Nature Genetics,* **40**, 768-775.

Meyre, D., Delplanque, J., Chevre, J-C., Lecoeur, C., Lobbens, S., Gallina, S., et al. (2009) Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nature Genetics,* **41**, 157-159.

Norton, E., Han, E. (2008) Genetic Information, Obesity and Labor Market Outcomes. *Health Economics*, **17**, 1089-1104.

Palmer, T.M., Lawlor, D.A., Harbord, R.M., Sheehan, N.A., Tobias, J.H., Timpson, N.J., Davey Smith, G., Sterne, J.A. (2011) Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods in Medical Research,* doi: 10.1177/0962280210394459.

Patsopoulos, N., Tatsioni, A., Ioannidis, J.P. (2007) Claims of Sex Differences: an Empirical Assessment in Genetic Associations. *JAMA,* **298**, 880-893.

Rubin, D.B. (1974) Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, **66**, 688-701.

Sheehan, N.A., et al. (2008) Mendelian Randomization and Causal Inference in Observational Epidemiology. *PLoS Medicine*, **5**, 1205-1210.

Shinohara, R.T. et al. (2009) Estimating Effects by Combining Instrumental Variables with Case-Control Designs: The Role of Principal Stratification. John Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 198.

Shinohara, R.T., Frangakis, C.E. (2009) The Role of Principal Stratification in Instrumental Variables in Case-Control Designs – An Application to Mendelian Randomization. Accessed on 1 August 2011 [Available at:] 193.206.143/molpage2009/ci-Frangakis.pdf

Speliotes, E.K., et al. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics,* **42**(11):937-950.

Thorleifsson, G., Walters, G.B., Gudbjartsson, D.F., Steinthorsdottir, V., Sulem, P., Helgadottir, A., et al. (2009) Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature Genetics,* **41,** 18-24.

Timpson, N.J., Emmett, P.M., Frayling, T.M., Rogers, I., Hattersley, A.T., McCarthy, M.I., Davey Smith, G. (2008) The Fat Mass- and Obesity-Associated Locus and Dietary Intake in Children. *The American Journal of Clinical Nutrition,* **88**, 971-978.

Timpson, N.J., Harbord, R., Davey Smith, G., Zacho, J., Tybjaerg-Hansen, A., Nordestgaard, B.G. (2009) Does Greater Adiposity Increase Blood Pressure and Hypertension Risk? Mendelian Randomization Using the FTO/MC4R Genotype. *Hypertension,* **54**, 84-90.

Wardle, J., Carnell, S., Haworth, C.M.A., Farooqi, S., O'Rahilly, S., Plomin, R. (2008) Obesity Associated Genetic Variation in *FTO* is associated with Diminished Satiety. *The Journal of Clinical Endocrinology and Metabolism,* **93**, 3640-3643.

Willer, C.J., Speliotes, E.K., Loos, R.J.F., Li, S., Lindgren, C.M., Heid, I.M., Berndt, S., et al. (2009) Six New Loci Associated with Body Mass Index Highlight a Neuronal Influence on Body Weight Regulation. *Nature Genetics,* **41**(1), 25-34.

Figures

Figure 1: Fat Mass Distribution Functions by the Instrumental Variable

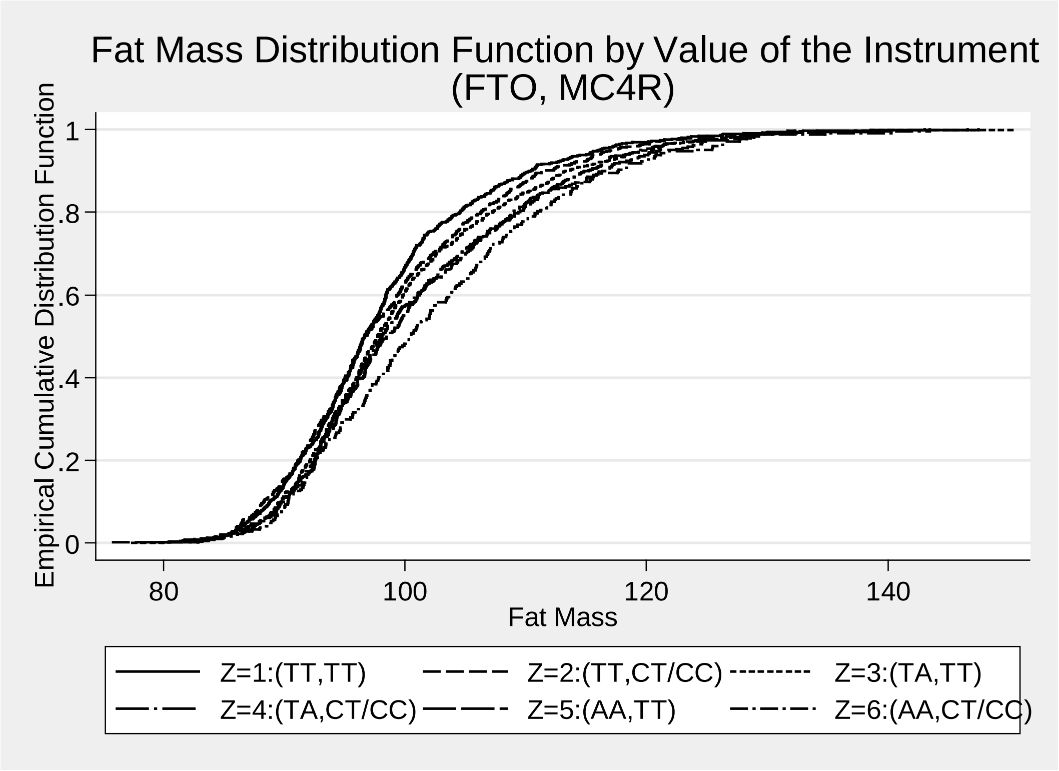
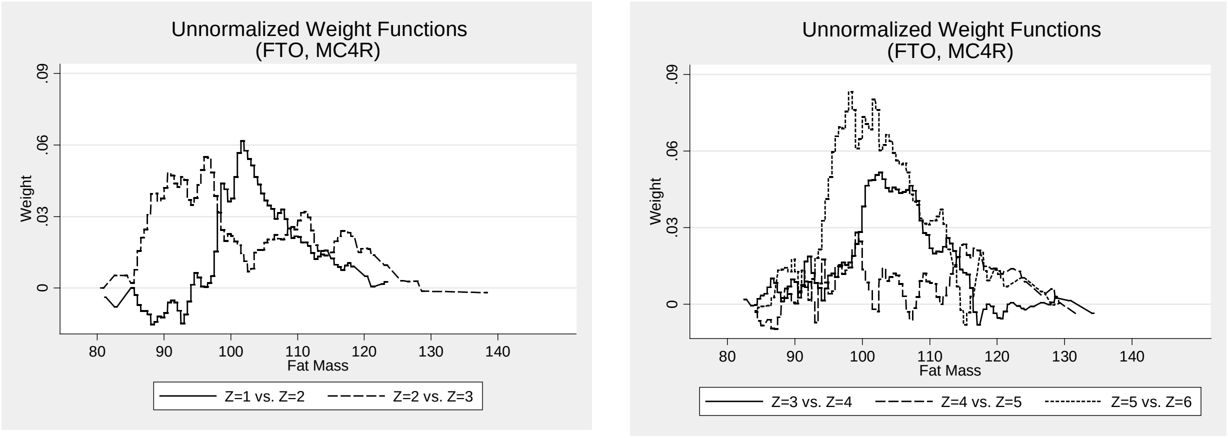


Figure 2: Unnormalised Weight Functions for the Different Instrument Values



Tables

Table 1: The six values of the instrumental variable

Z	*FTO*	*MC4R*	(*FTO*, *MC4R*)	Frequency	Mean fat mass
1	TT	TT	(TT,TT)	0.208	98.398
2	TT	CT/CC	(TT,CT/CC)	0.156	98.998
3	TA	TT	(TA,TT)	0.274	99.963
4	TA	CT/CC	(TA,CT/CC)	0.204	100.654
5	AA	TT	(AA,TT)	0.092	100.990
6	AA	CT/CC	(AA,CT/CC)	0.066	102.378

Table 2. OLS and IV estimates of the average response in standardised KS3

	(1) – OLS		(2) – First stage IV		(3) – Second stage IV	
Dependent variable:						
	KS3		Fat Mass		KS3	
	Coeff	*p-value*	*Coeff*	*p-value*	*Coeff*	*p-value*
Fat Mass						
Estimate and *p*-value	-0.099	*<0.001*			0.140	*0.282*
95% confidence interval	[-0.128, -0.070]				[-0.115, 0.396]	
Instrument						
Z=2: (TT,CT/CC)			0.600	*0.247*		
			[-0.416, 1.616]			
Z=3: (TA,TT)			1.564	*<0.001*		
			[0.687, 2.441]			
Z=4: (TA,CT/CC)			2.256	*<0.001*		
			[1.301, 3.210]			
Z=5: (AA,TT)			2.591	*<0.001*		
			[1.326, 3.857]			
Z=6: (AA,CT/CC)			3.980	*<0.001*		
			[2.477, 5.482]			
First stage *F*-statistic			9.365			
p-value, Hansen J-test					0.631	
p-value, Hausman test					0.053	
R-squared	0.01		0.014			
Number of children	3729		3729		3729	