American Journal of Epidemiology
Vol. 175, No. 4
DOI: 10.1093/aje/kwr323
Advance Access publication:
January 12, 2012


## Practice of Epidemiology

# Credible Mendelian Randomization Studies: Approaches for Evaluating the Instrumental Variable Assumptions

**M. Maria Glymour*, Eric J. Tchetgen Tchetgen, and James M. Robins**

* Correspondence to Dr. M. Maria Glymour, Department of Society, Human Development, and Health, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115 (e-mail: mglymour@hsph.harvard.edu).

*Initially submitted April 16, 2011; accepted for publication August 24, 2011.*

As with other instrumental variable (IV) analyses, Mendelian randomization (MR) studies rest on strong assumptions. These assumptions are not routinely systematically evaluated in MR applications, although such evaluation could add to the credibility of MR analyses. In this article, the authors present several methods that are useful for evaluating the validity of an MR study. They apply these methods to a recent MR study that used fat mass and obesity-associated (*FTO*) genotype as an IV to estimate the effect of obesity on mental disorder. These approaches to evaluating assumptions for valid IV analyses are not fail-safe, in that there are situations where the approaches might either fail to identify a biased IV or inappropriately suggest that a valid IV is biased. Therefore, the authors describe the assumptions upon which the IV assessments rely. The methods they describe are relevant to any IV analysis, regardless of whether it is based on a genetic IV or other possible sources of exogenous variation. Methods that assess the IV assumptions are generally not conclusive, but routinely applying such methods is nonetheless likely to improve the scientific contributions of MR studies.

causality; confounding factors; epidemiologic methods; instrumental variables; Mendelian randomization analysis

Abbreviations: BMI, body mass index; DAG, directed acyclic graph; IV, instrumental variable; MR, Mendelian randomization.

Mendelian randomization (MR) studies use genotypes as instrumental variables (IVs) to estimate the health effects of phenotypes influenced by those genotypes (1–6). MR-based effect estimates rest on strong assumptions (7–9), but MR applications often do not systematically evaluate these assumptions. Routinely presenting such evaluations would add to the credibility of MR studies (10–13). A recent article by Kivimäki et al. (14) stands out as an example in which the authors provide evidence regarding the plausibility of the MR assumptions. Kivimäki et al. used fat mass and obesity-associated (*FTO*) genotype as an IV to estimate the effect of body mass index (BMI; weight (kg)/height $(m)^2$) on risk of mental disorder. They found large and statistically significant IV effect estimates, suggesting that a high BMI increased the risk of mental disorder (though these findings have not been replicated (15)). Kivimäki et al. also provided results useful for evaluating the validity of the *FTO* genotype as an IV. They concluded that the findings suggest that *FTO* is not a valid IV and that the MR-based effect estimate is probably severely biased. Using Kivimäki et al.'s study as an example, we describe methods for evaluating the validity of an MR study (summarized in Table 1). To set the stage, we begin by reviewing the assumptions an IV must satisfy and causal structures that violate these assumptions.

## ASSUMPTIONS FOR MR STUDIES

The assumptions of MR studies can be represented using causal directed acyclic graphs (DAGs) (7, 8, 16–19). Figure 1A shows a set of assumptions under which *FTO* provides a valid IV for the effect of BMI on mental disorder:

1) *FTO* (the genotype) is associated with BMI (the phenotype);
2) there are no unmeasured common causes of *FTO* and mental disorder (the outcome); and
3) every directed pathway (sequence of arrows) from *FTO* to mental disorder passes through BMI.

**Table 1.** Four Alternative Empirical Approaches for Assessing the Validity of Proposed Instrumental Variables in Mendelian Randomization Studies

| Approach |
|---|
| Leverage prior causal assumptions regarding confounding of the phenotype-outcome association. There are 4 equivalent versions of this test, including the simple comparison of the Mendelian randomization effect estimate with a conventional effect estimate. The test relies on the assumption regarding the direction of confounding, and even with this assumption the test is not guaranteed to be consistent in at least some nonlinear causal structures. |
| Identify factors that modify the genotype-phenotype association. Compare the IV effect estimate across groups in which the population association between the instrument and the phenotype is either silenced or reversed. This test could identify a biased instrument provided that the biasing pathway is active in both subgroups. |
| Apply instrumental inequality tests. These tests are applicable only when the causal phenotype is known to be categorical. Even with categorical phenotypes, these inequality tests can only detect extreme violations of the assumptions, although the greater the number of IVs, the more sensitive the tests are. |
| Use multiple IVs to conduct overidentification tests. Other genes or even polymorphisms in the same gene might provide additional instruments. Overidentification tests cannot detect violations of the IV assumptions if all instruments have identical biasing pathways. They may also reject even when all instruments are valid if the model is incorrectly assumed to be linear or if the phenotype is composite. These tests generally have low statistical power. |

Abbreviation: IV, instrumental variable.

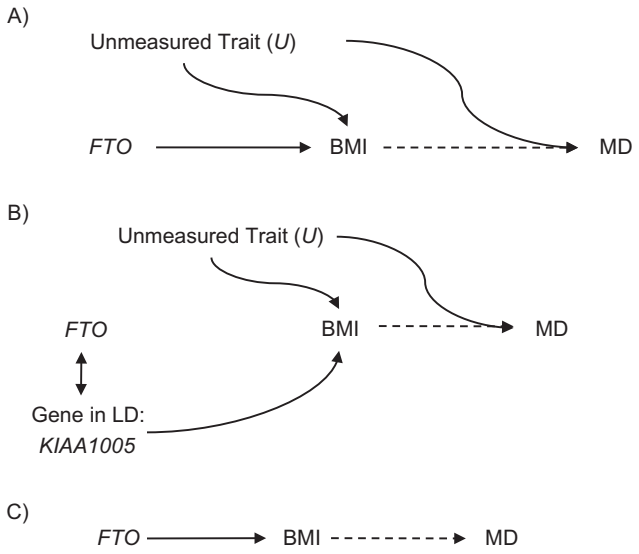


**Figure 1.** Causal structures in which fat mass and obesity-associated (*FTO*) genotype provides a valid instrumental variable (IV) with which to estimate the effect of body mass index (BMI) on the risk of mental disorder (MD). In these diagrams, dotted lines indicate the hypothesized causal pathway of primary research interest: the effect of BMI on MD. The solid arrows represent hypothesized causal pathways, and the absence of an arrow connecting 2 variables represents the assumption that these variables do not affect one another. Whenever 2 variables in the diagram share a common cause, that relation is shown in the diagram (even if the specific common cause is unknown or unmeasured (*U*)). Thus, the directed acyclic graph (DAG) in part A represents the assumptions that *FTO* and *U* both affect BMI, *U* affects MD, there is no direct effect of *FTO* on MD, and there is no common prior cause of *FTO* and MD. The DAG in part B introduces linkage disequilibrium (LD) between *FTO* and the *KIAA1005* gene. Under the assumptions shown in part B, *KIAA1005* would be a valid IV with which to estimate the effect of BMI on MD, and because there are no other pathways connecting *FTO* and MD, *FTO* is also a valid IV. The DAG in part C is similar to that in part A, but *U* has been eliminated; thus, part C represents the assumption that there are no confounders of BMI and MD.

*FTO* might not affect BMI but rather may be in linkage disequilibrium with an adjacent causal gene, for example, *KIAA1005* (20). This does not necessarily invalidate *FTO* as an IV, provided that *KIAA1005* fulfills the above assumptions and there are no other pathways linking *FTO* to mental disorder (as shown in Figure 1B, where linkage disequilibrium is represented as a double-headed arrow) (7).

It should be emphasized that *FTO* might be a valid IV with which to estimate the effect of BMI on mental disorder regardless of whether BMI affects mental disorder primarily via psychosocial mechanisms or biochemical mechanisms.

Among the 3 IV assumptions, assumption 1 is easily evaluated, and Kivimäki et al. show that *FTO* and BMI are positively associated (14). Assumptions 2 and 3 cannot be proven but can sometimes be falsified or shown to be inconsistent with prior evidence.

Parts A and B of Figure 1 show a vector *U* of unmeasured confounders of BMI and mental disorder, indicating that conventional analyses would be biased. The possibility of unmeasured confounding typically motivates MR studies. Kivimäki et al. (14) provide some evidence that the net residual confounding from unmeasured factors may be positive, by noting that associations between BMI and mental disorder have been less positive in studies that included statistical adjustment for many potential confounders. As is discussed below, such evidence suggests net positive confounding but is not conclusive. Nonetheless, in the remainder of this article we will assume unmeasured net positive confounding of the BMI-mental disorder association. Reasonable researchers may disagree with this assumption, but we do not further discuss its substantive merits. Our goal is to show how assumptions based on prior theory or evidence can be combined with empirical data to assess the validity of an instrument and to rule out certain causal structures. Without assumptions about the direction of the unmeasured confounding of the BMI-mental disorder association, there are fewer tools with which to evaluate the IV assumptions. Throughout, we suppose that the data have been stratified on all measured confounders of the BMI-mental disorder association that are unaffected by *FTO*.

In Figure 1C, *FTO* is a valid IV but the IV analysis is unnecessary because there is no unmeasured confounding.

The DAGs in Figure 2 show modifications of those in Figure 1, in which IV assumption 2 or 3 is violated and *FTO* is not a valid instrument. If IV assumption 2 or 3 is violated, the magnitude of bias in the standard IV effect estimate (defined below and in Web Appendix 1 (http://aje.oxfordjournals.org/)) is generally inflated in inverse proportion to the strength of the association between the instrument and the phenotype. Even very weak direct pathways from the gene to the outcome can

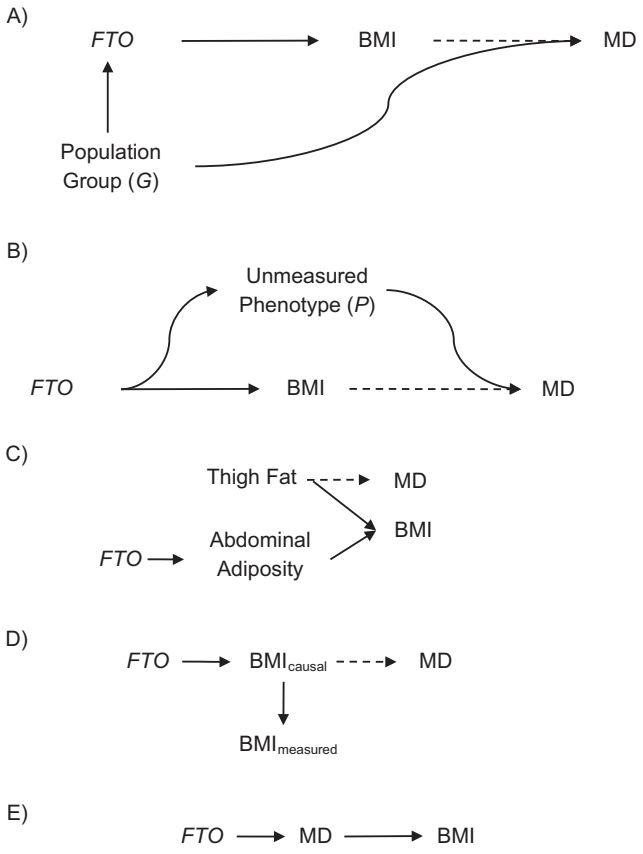


**Figure 2.** Causal structures in which fat mass and obesity-associated (*FTO*) genotype is not a valid instrumental variable (IV) with which to estimate the effect of body mass index (BMI) as measured on the risk of mental disorder (MD). *FTO* would not be a valid instrument for estimating the effect of BMI on MD if *FTO* and MD shared an unmeasured common cause such as population group (*G*), sometimes called population stratification (as in part A), or if there is a causal pathway from *FTO* to MD that is not mediated by BMI, as with pleiotropy (shown in part B). In part C, *FTO* influences abdominal adiposity but not thigh fat. Both thigh fat and abdominal adiposity influence BMI, but only thigh fat affects MD. The Mendelian randomization effect estimate based on *FTO* would not correspond to the effect of thigh fat on MD. If, as in the directed acyclic graph (DAG) shown in part D, BMI as measured is not the causal version of BMI, then the IV estimate based on measured BMI would correspond to the causal effect only under special circumstances. In the DAG shown in part E, MD affects BMI, rather than vice versa, and the IV estimate would not correspond to the effect of BMI on MD.

severely bias the IV effect estimate if the gene has a tiny effect on the phenotype.

In Figure 2A, the existence of a population group (*G*) that influences *FTO* and separately influences mental disorder violates assumption 2, a causal structure sometimes called ''population stratification'' (1). In Figure 2B, assumption 3 is violated by an (unmeasured) phenotype (*P*) that affects mental disorder and is affected by *FTO*, for example, due to pleiotropic effects of the gene (1). Figure 2C shows another concern: Alternative components or versions (e.g., abdominal obesity vs. thigh obesity) of the phenotype (BMI) may have different effects on the outcome. If so, MR can only identify the effect on the outcome of that version of the phenotype (abdominal obesity) which is induced by the genotype (*FTO*) (21). In Figure 2C, this effect, of abdominal obesity on mental disorder, is null. Absent knowledge of the relevant biology, this null effect might be misinterpreted as definitive evidence that no aspect of obesity affects mental disorder, overlooking the effect of thigh fat. This example shows how prior knowledge of the biologic, clinical, or social mechanisms linking the genotype and phenotype and the phenotype and the outcome can help assess the interpretation and plausibility of alternative causal structures (1, 2, 6).

Figure 2D shows a structure in which the causal BMI phenotype ($BMI_{causal}$) is measured with error and $BMI_{measured}$ has no causal effect on mental disorder. The error may represent any or all of the following: conventional random error, incorrect specification of the dose-response function linking BMI to mental disorder, or incorrect choice of the relevant exposure period. An example of the latter would be using adult BMI in the analysis when the causal phenotype is BMI in childhood. Because genetic factors can affect the phenotype throughout life, the choice of the relevant phenotype exposure period is of greater concern in MR than in many other IV applications.

Under the DAG shown in Figure 2D, *FTO* is a valid instrument for the effect of $BMI_{causal}$ on mental disorder, although $BMI_{measured}$ cannot necessarily be used to derive an unbiased MR effect estimate for $BMI_{causal}$. *FTO* is not a valid instrument for the (null) effect of $BMI_{measured}$ on mental disorder, because IV assumption 3 fails if the phenotype of interest is $BMI_{measured}$. Nonetheless, DAG 2D (unlike DAG 2B) implies that a test for an association between *FTO* and mental disorder remains a valid test of the null hypothesis of no effect of BMI on mental disorder (i.e., of the hypothesis that the arrow from $BMI_{causal}$ to mental disorder is absent).

In Web Appendix 1, we show that if $BMI_{causal}$ affects risk of mental disorder, the MR estimate using $BMI_{measured}$ may be unbiased, inflated, or attenuated compared with the estimate based on $BMI_{causal}$, depending on the regression slope of $BMI_{measured}$ on $BMI_{causal}$ (i.e., the nature of the mismeasurement). Under the classical measurement error model ($BMI_{measured}$ equals $BMI_{causal}$ plus mean zero independent random error), the MR estimand based on $BMI_{measured}$ equals that based on $BMI_{causal}$. In contrast, suppose $BMI_{measured}$ and $BMI_{causal}$ were dichotomous and misclassification were nondifferential—that is, the misclassification probabilities did not depend on *FTO* or mental disorder. In this situation, the magnitude of the MR estimate based on $BMI_{measured}$ would *exceed* that based on $BMI_{causal}$. However, the MR estimate based on $BMI_{measured}$ will be *smaller* than that based on $BMI_{causal}$, if, when $BMI_{measured}$ is regressed on $BMI_{causal}$, the slope of the regression line exceeds 1. This may be common in MR studies and would apply, for example, if $BMI_{causal}$ was BMI in childhood, $BMI_{measured}$ was adult BMI, or $BMI_{adult} = 1.2 \times BMI_{childhood}$ plus an independent measurement error.

In Figure 2E, IV assumption 3 is violated because *FTO* affects risk of mental disorder, which in turn affects BMI. This elaboration of Figure 2B is especially relevant to MR studies because the temporal order of the phenotype and outcome measurements often provides little information, or

even misleading information, regarding the causal direction linking the phenotype and the outcome.

We next turn to some empirical approaches to evaluating whether proposed IVs are valid (summarized in Table 1).

## FALSIFYING IV ASSUMPTIONS BY LEVERAGING PRIOR CAUSAL ASSUMPTIONS

One approach to testing IV assumptions is to attempt to determine whether the data are compatible with prior assumptions of positive residual confounding of the phenotype-outcome association. We first explain the approach and then provide informal intuitions; relevant proofs are given in Web Appendix 2. We show that under the assumption that net unmeasured confounding is positive, this approach often (but not always) provides a valid test of the IV assumptions.

Constructing the test requires a formal definition of positive net unmeasured confounding. We provide a useful counterfactual definition that allows for the possibility that the vector $U$ of common causes of mental disorder and BMI contains components that result in negative confounding, provided that *net* unmeasured confounding is positive. Let $FTO = 0, 1, 2$ denote the number of minor *FTO* alleles a particular individual carries. Define the counterfactual BMI($FTO = z$) to be a subject's (centered) BMI if, possibly contrary to fact, he was of type $z$ at the *FTO* locus. **BMI(*FTO*)** represents the set (bolded to indicate a set) of counterfactual BMI values for this subject for all possible values of *FTO*. Let MD(BMI $= x$) be a subject's counterfactual mental disorder (MD) if, possibly contrary to fact, his BMI was $x$. **MD(BMI)** represents the set of counterfactual values of mental disorder for all possible values of BMI. This notation incorporates the IV assumption that mental disorder depends on BMI but not further on *FTO*.

*Definition*: Positive net confounding of the association between mental disorder and BMI exists if MD(BMI $= x$) and BMI($FTO = z$) are positively correlated for all possible values $x$ of BMI and $z$ of *FTO*.

This is consistent with conventional accounts of confounding. It holds if BMI and mental disorder share one common cause which positively affects both conditions or multiple common causes, some of which might result in negative confounding, provided the magnitude of positive confounding exceeds the negative confounding. Under our definition, adjusting for positive confounders would generally reduce the magnitude of the coefficient for BMI in the regression of mental disorder on BMI, consistent with prior studies of BMI and mental disorder noted in Kivimäki et al.'s paper (14).

This definition is motivated by Figure 3, which adds **MD(BMI)** and **BMI(*FTO*)** to DAG 1A. This DAG encodes the fact that $U$ influences the observed BMI and mental disorder only through its effects on the counterfactuals **BMI(*FTO*)** and **MD(BMI)**. If the unmeasured common causes $U$ positively (or negatively) affect both **BMI(*FTO*)** and **MD(BMI)**, then the counterfactuals will be positively correlated.

Assume for now a linear causal model in which the effect of BMI on mental disorder is linear with the same slope $r$ for every subject, so that MD(BMI $= x$) $= r \times x +$ MD(BMI $= 0$). For any 3 random variables ($X$, $Y$, $Z$), let $\boldsymbol{C_{XY}}$ denote the coefficient of $X$ in the ordinary least squares regression of

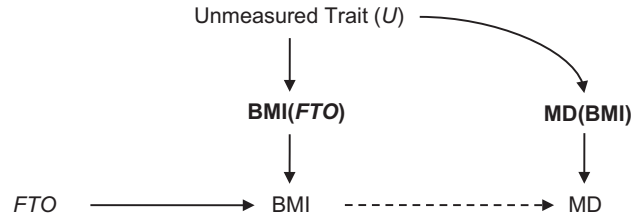


**Figure 3.** Directed acyclic graph showing that positive confounding arises from the association between counterfactuals. **BMI(*FTO*)** represents the set of counterfactual values of body mass index (BMI) for all possible values of fat mass and obesity-associated (*FTO*) genotype. **MD(BMI)** represents the set of counterfactual values of mental disorder (MD) for all possible values of BMI. An unmeasured trait ($U$) does not affect *FTO*, but $U$ does affect the value BMI would take for any specific value of *FTO*.

$Y$ on (1, $X$) and $\boldsymbol{C_{XY|Z}}$ correspond to the coefficient of $X$ in the regression of $Y$ on (1, $X$, $Z$). In the regression of mental disorder on *FTO*, the coefficient for *FTO* would be represented by $\boldsymbol{C_{FTO,MD}}$. When *FTO* is a valid instrument, the causal slope $r$ describing the effect of BMI on mental disorder equals the standard IV estimand: the ratio of $\boldsymbol{C_{FTO,MD}}$ to $\boldsymbol{C_{FTO,BMI}}$. The standard IV estimate referred to above is the sample version of this ratio.

We now describe the tests of positive unmeasured confounding. Given that *FTO* and BMI are positively correlated, we prove in Web Appendix 2 that if *FTO* is a valid IV and the linear causal model holds, then positive unmeasured confounding of BMI and mental disorder implies the following 4 equivalent statements.

1. The IV effect estimate is less than the ordinary least squares effect estimate: $\boldsymbol{C_{FTO,MD}/C_{FTO,BMI} < C_{BMI,MD}}$.
2. In the regression predicting mental disorder with *FTO* and BMI, the coefficient for *FTO* is negative: $\boldsymbol{C_{FTO,MD|BMI} < 0}$.
3. The estimated slope for BMI predicting mental disorder is closer to the IV estimate than is the slope for BMI predicting mental disorder with additional adjustment for *FTO*: $|\boldsymbol{C_{BMI,MD}\text{-}C_{FTO,MD}/C_{FTO,BMI}}| < |\boldsymbol{C_{BMI,MD|FTO}\text{-}C_{FTO,MD}/C_{FTO,BMI}}|$. Under the linear causal model, this implies that the magnitude of the bias of the ordinary least squares estimate for the causal slope $r$ adjusted for *FTO* exceeds the bias of the ordinary least squares estimate without adjustment for *FTO*.
4. The residual MD $-$ ($\boldsymbol{C_{FTO,MD}/C_{FTO,BMI}}$) $\times$ BMI is positively correlated with BMI: Cov(MD $-$ ($\boldsymbol{C_{FTO,MD}/C_{FTO,BMI}}$) $\times$ BMI, BMI) $> 0$. Under the linear causal model, this residual is an estimate of the counterfactual MD(BMI $= 0$).

If confounding is positive and the linear causal model holds, then empirical violations of any of the above 4 statements imply that *FTO* is not a valid instrument. More generally, this result is true if a structural nested mean model with no effect modification or treatment interaction holds (18). A linear causal model is a special case.

Kivimäki et al. (14) found that the IV estimate $\boldsymbol{C_{FTO,MD}/C_{FTO,obesity}}$ was more than 10-fold larger than $\boldsymbol{C_{obesity,MD}}$

(0.907 vs. 0.064), violating statement 1. Kivimäki et al. reported an unadjusted coefficient regressing mental disorder score on *FTO* ($\boldsymbol{C_{FTO,MD}}$) of 0.074 (95% confidence interval: 0.019, 0.129) among men (14). In the model including both *FTO* and BMI as independent variables, Kivimäki et al. found that the adjusted *FTO* coefficient ($\boldsymbol{C_{FTO,MD|BMI}}$) was 0.071 ($P = 0.006$ for the 1-sided test of the null hypothesis that the adjusted estimate was $\leq 0$), violating statement 2. Kivimäki et al. did not provide an estimate of $\boldsymbol{C_{BMI,MD|FTO}}$ or $\text{Cov}(\text{MD} - (\boldsymbol{C_{FTO,MD}}/\boldsymbol{C_{FTO,BMI}}) \times \text{BMI}, \text{BMI})$ and thus did not directly test statement 3 or 4. However, all 4 statements are logically equivalent, so violations of statements 1 and 2 imply that statements 3 and 4 do not hold.

Statements 1–4 describe relations between empirical regression coefficients. The logical equivalence between these statements is a statistical fact that is true for any 3 measured variables (this is not generally recognized; see Web Appendix 2 for the proof), regardless of whether 1) the assumptions under which *FTO* is a valid instrument hold, 2) the BMI-mental disorder relation is positively confounded, or 3) the linear causal model is true. The only connection to causality is that $\boldsymbol{C_{FTO,MD}}/\boldsymbol{C_{FTO,BMI}}$ happens to be the causal coefficient $r$ if the linear causal model is correctly specified and *FTO* is a valid IV.

The informal intuition for why violation of statement 1 appears to imply that *FTO* is not a valid instrument is that if the conventional estimate is positively biased, a valid IV estimate should be smaller than the conventional effect estimate. However, if *FTO* is not a valid IV, there exists an open path from *FTO* to mental disorder that does not pass through BMI (as in part B or C of Figure 2). This path increases the *FTO*-mental disorder association $\boldsymbol{C_{FTO,MD}}$, without increasing the *FTO*-BMI association $\boldsymbol{C_{FTO,BMI}}$. Thus, the ratio $\boldsymbol{C_{FTO,MD}}/\boldsymbol{C_{FTO,BMI}}$ will be inflated by this open path. The weaker the association between the genotype and phenotype ($\boldsymbol{C_{FTO,BMI}}$), the greater the degree of inflation. This explanation applies regardless of whether the bias in the numerator is attributable to violations of IV assumption 2 or 3. Assessing statement 1 is only informative regarding the validity of the IV if we have a strong prior assumption about the direction of confounding of the conventional effect estimate. If the unmeasured confounding of the BMI-mental disorder association could plausibly be negative, then finding that the IV estimate is much larger than the conventional estimate does not suggest that *FTO* is an invalid instrument.

Applications of the logically equivalent test in statement 2 have been used in prior MR analyses (11). To see the intuition for statement 2, note that in Figure 1A (showing a valid IV), adjusting for BMI would render *FTO* and mental disorder statistically independent except for the association induced by "collider bias." Collider bias arises because BMI is a common effect of *FTO* and the unmeasured confounder $U$; conditional upon BMI, *FTO* and $U$ become statistically associated (22). In Web Appendix 2 we show that, in all linear models and many nonlinear models, this collider bias will be negative provided that the net unmeasured confounding of the BMI-mental disorder association induced by $U$ is positive. Because with a valid IV this collider bias is the only source of statistical association between BMI and mental disorder conditional upon BMI, the regression coefficient $\boldsymbol{C_{FTO,MD|BMI}}$ will likewise be negative. Statement 3 is essentially equivalent to Wooldridge's (23) and Pearl's (24) recent results showing that if true effects are additive, adjusting for valid IVs exacerbates bias in associations estimated with ordinary least squares.

Despite these intuitions, these criteria cannot be used to conclusively disprove the validity of the IV. Even when *FTO* is a valid IV and confounding is positive, we show in Appendix 2 that there exist certain extremely nonlinear causal models for which statements 1–4 are false. However, we also show that other, quite nonlinear causal models agree with our result for linear models: Under positive confounding, if statements 1–4 are false, then *FTO* is not a valid IV.

## FALSIFYING THE IV ASSUMPTIONS BY IDENTIFYING MODIFYING SUBGROUPS

A second approach sometimes used in efforts to falsify the MR assumptions is to identify a subgroup among whom the genotype does not predict the phenotype (13, 25, 26). For example, in Kivimäki et al.'s results (14), *FTO* is independent of BMI among women. Figure 4A is consistent with this structure: The only path linking *FTO* and BMI is mediated

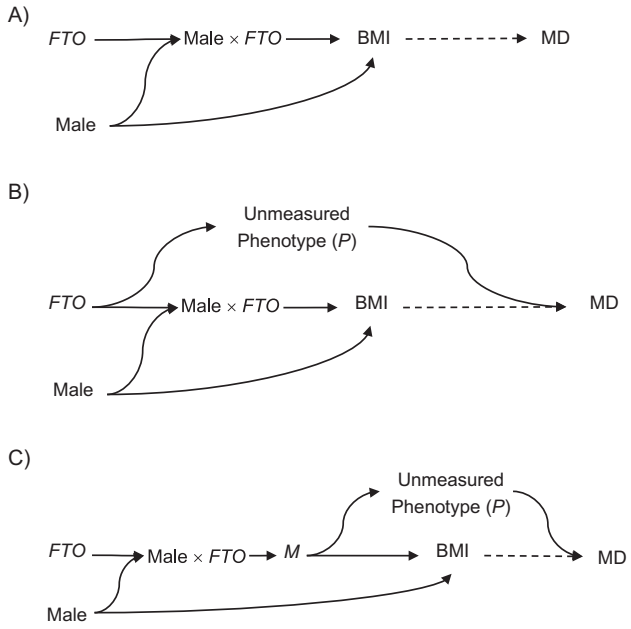


**Figure 4.** Causal structure in which the effect of fat mass and obesity-associated (*FTO*) genotype on body mass index (BMI) is silenced among women. In part A, *FTO* is a valid instrument and the association between *FTO* and mental disorder (MD) is silenced among women because the variable Male × *FTO* is zero among women (modified from Figure 1C). In part B, *FTO* is not a valid instrument, and the biasing pathway linking *FTO* and MD would create a statistical association between *FTO* and MD even among women (this is a modification of Figure 2B). In part C, *FTO* is not a valid instrument, but the biasing pathway linking *FTO* and MD would not create a statistical association between *FTO* and MD even among women, because the biasing pathway with the unmeasured phenotype originates after the Male × *FTO* interaction. In this diagram, the variable *M* represents a mediator which is influenced by Male × *FTO* interaction and affects both BMI and the unmeasured phenotype (this is also a modification of Figure 2B).

by the interaction of male gender and *FTO* (27). Among women, this interaction variable is always zero (i.e., for women, male $\times$ *FTO* $= 0$): The path is silenced. Because this path is silenced, *FTO* is statistically independent of BMI and mental disorder among females if there are no other pathways linking *FTO* to mental disorder. As shown in Figure 4B, if *FTO* has a pleiotropic pathway to mental disorder which is not mediated by this interaction, the biasing pathway will create a statistical association between *FTO* and mental disorder even among women. Thus, if *FTO* predicts mental disorder in the subgroup (women) in which *FTO* does not predict BMI, this suggests that the MR assumptions are violated. Kivimäki et al. find that *FTO* is independent of mental disorder among women, consistent with assumptions under which *FTO* would provide a valid instrument. Again, this is not conclusive evidence supporting the validity of the instrument. An invalid instrument may "pass" this test, if both the BMI-mediated pathway and the biasing phenotype have a common cause, say *M*, subsequent to *FTO* and thus subsequent to the male $\times$ *FTO* interaction (Figure 4C). In this case, the biasing phenotype is only active among men.

Using "modifying subgroups" to assess the IV assumptions is likely to be most convincing in situations where reasons for subgroup differences in the genotype-phenotype association are well understood. The association between *FTO* and BMI has been demonstrated in other female samples, suggesting that the independence observed among women in the Whitehall II study (14) may be a statistical fluke.

## FALSIFYING THE IV ASSUMPTIONS WITH INSTRUMENTAL INEQUALITY TESTS

A third approach to testing for violations of IV assumptions, applying Bonet's "instrumental inequality tests" (19, 28), can be used with categorical phenotypes. These tests only detect rather extreme violations of the IV assumptions. To illustrate, suppose we hypothesize a risk threshold at a particular BMI cutpoint (obesity) and that BMI levels above or below this cutpoint do not further affect the risk of mental disorder. The IV assumptions then imply that a valid IV satisfies the following inequalities (where *Y* is an indicator for 1 or more episodes of mental disorder; the binary variable *X* represents the presence or absence of obesity; and *Z* represents *FTO* minor allele frequency, taking the value 0, 1, or 2):

- $\max_i[\Pr(X = 0, Y = 1 \mid Z = i)]$
  $\leq \min_i[1 - \Pr(Y = 0, X = 0 \mid Z = i)]$.
- $\max_i[\Pr(X = 1, Y = 1 \mid Z = i)]$
  $\leq \min_i[1 - \Pr(Y = 0, X = 1 \mid Z = i)]$.
- $\max_i[\Pr(X = 0, Y = 1 \mid Z = i)]$
  $+ \Pr(X = 1, Y = 1 \mid Z = i)]$
  $+ \max_i[\Pr(X = 0, Y = 1 \mid Z = i)]$
  $+ \Pr(X = 1, Y = 0 \mid Z = i)]$
  $+ \max_i[\Pr(X = 0, Y = 0 \mid Z = i)] \leq 2$.

Based on data provided to us by Kivimäki et al. (14), we confirmed that all of these inequalities were satisfied, so we could not reject *FTO* as a valid IV using these tests (Web Appendix 3). If BMI levels above or below the cutpoint independently affect mental disorder, then the IV assumptions for dichotomized obesity are certainly violated, even if the Bonet tests do not reject them.

## FALSIFYING THE IV ASSUMPTIONS WITH OVERIDENTIFICATION TESTS

Overidentification tests are another promising approach to evaluating IVs (29), but they require multiple IVs. Overidentification tests use Sargan-type statistics to evaluate the null hypothesis that effect estimates from multiple IVs are identical: If they differ significantly, intuition suggests that at least 1 of the putative instruments is not valid (29, 30). If the estimates from multiple instruments are all similar and thus the overidentification test does not reject them, intuition suggests that none are biased. However, overidentification tests cannot rule out the possibility that all of the IVs are biased in the same way. Overidentification tests can use multiple genes, each of which influences the phenotype of interest; they will thus become increasingly feasible as MR studies use genome-wide data (10). It is also possible to use multiple polymorphisms in the same gene or heterozygote/homozygote contrasts as instruments (31), although this approach is somewhat less reassuring when the instruments appear to "pass" the overidentification test. The value of overidentification tests rests in our intuition that biasing pathways (e.g., DAG 2A or 2B) could not plausibly be identical for different IVs. When alternative instruments are based on multiple polymorphisms of 1 gene, this intuition is less appealing. It seems possible that polymorphisms of the same gene could all trigger identical biasing pathways, giving identically biased IV effect estimates.

There are important limitations of overidentification tests, regardless of the source of the different instruments. They can neither conclusively verify nor conclusively falsify proposed instruments. Even when all of the instruments are valid, overidentification tests may inappropriately reject them if the phenotype-outcome association is falsely assumed to be linear. Overidentification tests may also inappropriately reject if all IVs are valid but affect different components or versions of the phenotype which themselves have differing effects on the outcome.

Furthermore, overidentification tests may inappropriately fail to reject even when one or more of the putative instruments are invalid, because the tests usually have low statistical power (10, 29). IV effect estimates tend to be imprecise, so an overidentification test of whether 2 IV estimates are consistent with each other may not reject even when one or both are severely biased. With multiple genetic IVs, it may be possible to combine instruments to develop more precise estimates and stronger overidentification tests (10, 32, 33).

## CONCLUSION

It is not generally possible to prove MR assumptions, but it is often possible to find empirical (though usually not conclusive) evidence suggesting that the putative IV is invalid.

There is no fail-safe or certain approach to evaluating proposed IVs: Any of the proposed tools for evaluating IV assumptions may fail to identify biased IVs under various circumstances. This does not preclude the potential contributions of IV or MR methods. Rather, it implies that standards long accepted in observational epidemiology should extend to IV results. Understanding of causal effects is generally advanced by triangulation of evidence from multiple alternative sources, preferably sources which do not all rest on identical assumptions; transparency of analytic assumptions; and concerted efforts to verify or falsify those assumptions. Given that it will often be nearly free to conduct IV analyses with secondary data, they may prove extremely valuable in many research areas. Evidence from MR studies will be most appealing when 1) genotype-phenotype associations are both strong and physiologically comprehensible or multiple independent genotypes provide similar MR effect estimates and 2) the MR study is replicable in different populations. Potential contributions of MR are greatest when there is good reason to worry that conventional studies are biased; in this case, even imperfect evidence from MR studies might strengthen the evidence base. We recommend that when employing MR applications, investigators routinely use applicable methods to evaluate the IV assumptions. Although IV estimates are often quite sensitive to violations of assumptions, sensitivity analyses illustrating whether plausible violations could account for observed IV effect estimates would be valuable areas for future IV research. Even when MR assumptions are violated, critical evaluation may provide information regarding the pathways linking the genotype and phenotype (26). In Web Appendix 4, we discuss the causal structure most consistent with the evidence Kivimäki et al. present.

The accumulation of additional genetic information will probably expand the range of feasible MR studies (34, 35). If MR evolves into a reliable tool for causal discovery, the significance for health research could be great. On the other hand, if MR is uncritically adopted into the epidemiologic toolbox, without aggressive evaluations of the validity of the design in each case, it may generate a host of false or misleading findings.

## ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts (Eric J. Tchetgen Tchetgen, James M. Robins); Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts (Eric J. Tchetgen Tchetgen, James M. Robins); and Department of Society, Human Development, and Health, Harvard School of Public Health, Boston, Massachusetts (M. Maria Glymour).

The authors gratefully acknowledge financial support from the National Institute of Mental Health (grant MH092707-01) and the National Institute of Environmental Health Sciences (grant ES019712-01) and helpful comments on an earlier draft of this article by Dr. David Rehkopf.

Conflict of interest: none declared.

## REFERENCES

1. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol.* 2003;32(1):1–22.
2. Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol.* 2004;33(1):30–42.
3. Davey Smith G, Ebrahim S. What can mendelian randomisation tell us about modifiable behavioural and environmental exposures? *BMJ.* 2005;330(7499):1076–1079.
4. Davey Smith G, Ebrahim S, Lewis S, et al. Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet*. 2005;366(9495):1484–1498.
5. Sheehan NA, Didelez V, Burton PR, et al. Mendelian randomisation and causal inference in observational epidemiology. *PloS Med.* 2008;5(8):e177. (doi:10.1371/journal.pmed.0050177).
6. Lawlor DA, Harbord RM, Sterne JA, et al. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med.* 2008;27(8):1133–1163.
7. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res.* 2007;16(4):309–330.
8. Didelez V, Meng S, Sheehan NA. Assumptions of IV methods for observational epidemiology. *Stat Sci.* 2010;25(1):22–40.
9. Sheehan NA, Meng S, Didelez V. Mendelian randomisation: a tool for assessing causality in observational epidemiology. *Methods Mol Biol.* 2011;713(4):153–166.
10. Palmer TM, Lawlor DA, Harbord RM, et al. Using multiple genetic variants as instrumental variables for modifiable risk factors [published online ahead of print January 7, 2011]. *Stat Methods Med Res.* (doi:10.1177/0962280210394459).
11. Timpson NJ, Sayers A, Davey-Smith G, et al. How does body fat influence bone mass in childhood? A Mendelian randomization approach. *J Bone Miner Res.* 2009;24(3):522–533.
12. Timpson NJ, Harbord R, Davey Smith G, et al. Does greater adiposity increase blood pressure and hypertension risk?: Mendelian randomization using the *FTO/MC4R* genotype. *Hypertension*. 2009;54(1):84–90.
13. Chen L, Davey Smith G, Harbord RM, et al. Alcohol intake and blood pressure: a systematic review implementing a Mendelian randomization approach. *PloS Med*. 2008;5(3):e52. (doi:10.1371/journal.pmed.0050052).
14. Kivimäki M, Jokela M, Hamer M, et al. Examining overweight and obesity as risk factors for common mental disorders using fat mass and obesity-associated (*FTO*) genotype-instrumented analysis: the Whitehall II Study, 1985–2004. *Am J Epidemiol.* 2011;173(4):421–429.
15. Lawlor DA, Harbord RM, Tybjaerg-Hansen A, et al. Using genetic loci to understand the relationship between adiposity and psychological distress: a Mendelian randomization study in the Copenhagen General Population Study of 53,221 adults. *J Intern Med.* 2011;269(5):525–537.
16. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology.* 1999;10(1):37–48.
17. Glymour MM, Greenland S. Causal diagrams. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology.* 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008:183–210.
18. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology.* 2006;17(4):360–372.
19. Pearl J. *Causality.* Cambridge, United Kingdom: Cambridge University Press; 2000.

20. Frayling TM, Timpson NJ, Weedon MN, et al. A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*. 2007; 316(5826):889–894.
21. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc.* 1996; 91(434):444–455.
22. Cole SR, Hernán MA. Fallibility in estimating direct effects. *Int J Epidemiol.* 2002;31(1):163–165.
23. Wooldridge JM. *Should Instrumental Variables Be Used as Matching Variables?* (Technical report). East Lansing, MI: Michigan State University; 2009. (https://www.msu.edu/~ec/faculty/wooldridge/current%20research/treat1r6.pdf). (Accessed June 20, 2010).
24. Pearl J. On a class of bias-amplifying variables that endanger effect estimates. In: Grünwald P, Spirtes P, eds. *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*. Corvallis, OR: AUAI Press; 2010:417–424.
25. Davey Smith G. Mendelian randomization for strengthening causal inference in observational studies: application to gene by environment interaction. *Perspect Psychol Sci.* 2010;5(5): 527–545.
26. Glymour M, Veling W, Susser E. Integrating knowledge of genetic and environmental pathways to complete the developmental map. In: Kendler K, Jaffee S, Romer D, eds. *The Dynamic Genome and Mental Health: The Role of Genes and Environments in Youth Development*. New York, NY: Oxford University Press; 2011:172–194.
27. VanderWeele TJ, Robins JM. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *Am J Epidemiol.* 2007;166(9):1096–1104.
28. Bonet B. Instrumentality tests revisited. In: Breese JS, Koller D, eds. *UAI '01: Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann Publishers, Inc; 2001:48–55.
29. Hausman JA. Specification tests in econometrics. *Econometrica.* 1978;46(6):1251–1271.
30. Sargan J. The estimation of economic relationships using instrumental variables. *Econometrica*. 1958;26(3):393–415.
31. Ding EL, Song Y, Manson JE, et al. Sex hormone-binding globulin and risk of type 2 diabetes in women and men. *N Engl J Med.* 2009;361(12):1152–1163.
32. Pierce BL, Ahsan H, VanderWeele TJ. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Int J Epidemiol.* 2011;40(3):740–752.
33. Davey Smith G. Random allocation in observational data: how small but robust effects could facilitate hypothesis-free causal inference. *Epidemiology.* 2011;22(4):460–463.
34. Linsel-Nitschke P, Götz A, Erdmann J, et al. Lifelong reduction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of coronary artery disease—a Mendelian randomisation study. *PLoS ONE*. 2008;3(8):e2986. (doi:10.1371/journal.pone.0002986).
35. Shah SH, de Lemos JA. Biomarkers and cardiovascular disease: determining causality and quantifying contribution to risk assessment. *JAMA*. 2009;302(1):92–93.