Mendelian randomization (MR)

Use inherited genetic variants to infer causal relationship of an exposure and a disease outcome.

1. Concepts of MR and Instrumental variable (IV) methods
   - motivation, assumptions, inference goals, merits and limitations
   - two-stage least squares (2SLS) method from econometrics literature
   - Sargan's test for validity of IV
   - Durbin-Wu-Hausman test for equality of IV and OLS

2. Development of MR methods for binary disease outcomes
   - Various approximation methods extended from (2SLS)
   - Potential outcomes, structural mean models, consistent estimation of causal odds ratio
   - Model diagnostics

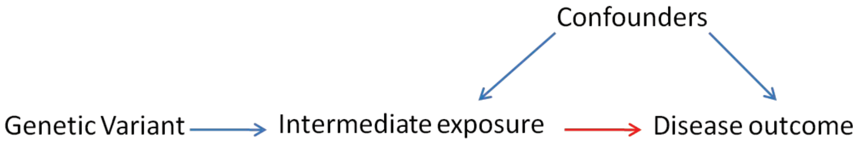
Hill's criteria: when can observed association be interpreted as causal?

Association $\neq$ Causality

- Strength: Lung cancer death rate in smokers about 9-10 times as non-smokers
- Consistency: repeatedly observed
- Specificity: certain type of disease but not others
- Temporality: cause precedes consequence
- Biological gradient: dose response, for example, lung cancer risk rises linearly with #cigarettes smoked daily
- Biological plausible
- Coherent with lab evidence
- Experimental or semi-experimental: if exposure was remove, does that prevent the disease?
- Analogy with known exposure-disease causal effect

Hill AB. Proceedings of the Royal Society of Medicine. 1965

Mendelian randomization analysis



The fundamental idea: If we cannot randomize the exposure, we can find a randomized instrumental variable to disentangle

- Confounding
- Reverse causation

Part I

Mendelian randomization: concepts, assumptions, 2SLS, etc

Katan M. Lancet 1986: a one-page letter

- Low cholesterol levels are sometimes associated with increased cancer risks, but it could be reverse causation.
- Differences in the amino acid sequence of apolipoprotein E (apo E) are major determinants of plasma cholesterol levels: E-2, E-3, E-4 with increased cholesterol levels.
- "if a naturally low cholesterol favours tumour growth, then subjects with the E-2/E-2 or E-2/E-3 phenotype should have an increased risk of cancer."

Reverse causation

"Unlike most other indices of lipid metabolism, apolipoprotein amino acid sequences are not disturbed by disease, and the apo E phenotype found in a patient will have been present since birth."

The reasoning behind Katan M. Lancet 1986

$$\text{low cholesterol} \Longleftrightarrow \text{increased cancer risk}$$

1 Apo E sequence variation $\Rightarrow$ low cholesterol, this relationship is established since inheritance of Apo E sequence variation

2 Apo E sequence variation $\Rightarrow$ increase cancer risk, cancer occurs later stage of life

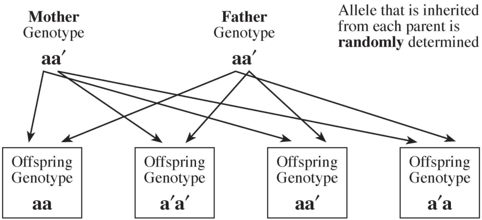
Two underlying assumptions: Apo E genetic effect on cancer risk can be unbiased assessed; Apo E genetic variation does not increase cancer risk through other pathways.

George Davey Smith and Shah Ebrahim 2003 IJE: first expanded presentation of MR

Confounding

"One key point is that the distribution of such polymorphisms is largely unrelated to the sorts of confounderssocioeconomic or behaviouralthat were identified above as having distorted interpretations of findings from observational epidemiological studies."

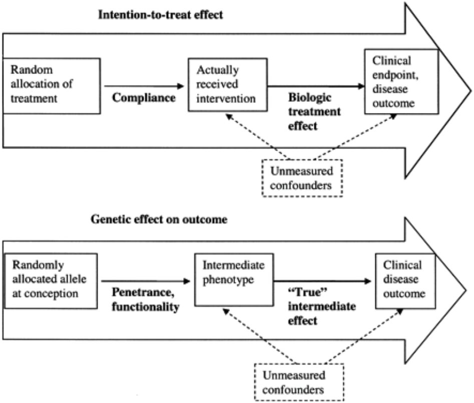
- Mendel's second law, the law of independent assortment, germline genetic variants can be viewed as if "randomized" conditional on parental genotypes.



Mother
Genotype
aa′
Father
Genotype
aa′
Allele that is inherited from each parent is **randomly** determined
Offspring Genotype **aa**
Offspring Genotype **a′a′**
Offspring Genotype **aa′**
Offspring Genotype **a′a**

- But it is an approximation in the population!

Conceptual analogy between MR and randomized clinical trials (RCT)


Intention-to-treat effect
Random allocation of treatment
Compliance
Actually received intervention
Biologic treatment effect
Clinical endpoint, disease outcome
Unmeasured confounders
Genetic effect on outcome
Randomly allocated allele at conception
Penetrance, functionality
Intermediate phenotype
"True" intermediate effect
Clinical disease outcome
Unmeasured confounders

- In RCT, confounding is removed by strict randomization. MR has at best "approximate randomization".
- In RCT, assignment exerts effect on disease endpoints through actually treatment received. MR has to assume that there is no direct effect from gene to disease (no other pathway).

Nitsch et al. Am J Epidemiol 2006

The objectives of MR studies

Analytical goals with proper assumptions

- Testing causal relationship between intermediate phenotype and disease outcome, by testing association between genotypic instrument and disease outcome: hypothesis testing - the Katan's original reasoning
- In linear and sometimes logistic models, estimating causal effect of intermediate on disease outcome: effect estimation after connecting to instrumental variables approach in econ

Three core assumptions for hypothesis testing

G: genetic variant; Y: disease outcome; X: intermediate exposure;
U: unknown confounder

Core assumptions

1 independence between G and U, (covariate adjustment)

$$G \perp U$$

2 established association between G and X, (strong/weak instrument)

$$\Pr(X|G) \neq \Pr(X)$$

3 no alternative pathway from G to Y, (exclusion restriction)

$$G \perp Y|X, U$$

$$\Downarrow$$

Testing the $G - Y$ association is equivalent to testing causal relationship $Y - X$.

Testing causal relationship

How is this derived mathematically?

$$\begin{aligned}\Pr(Y,G) &= \Pr(G)\int_u \Pr(U|G)\int_x \Pr(Y|G,X,U)\Pr(X|G,U)\\ &= \Pr(G)\int_u \Pr(U)\int_x \Pr(Y|X,U)\Pr(X|G,U)\end{aligned}$$

If $Y \perp X|U$, i.e., $\Pr(Y|X,U) = \Pr(Y|U)$,

$$\begin{aligned}\Pr(Y,G) &= \Pr(G)\int_u \Pr(U)\Pr(Y|U)\int_x \Pr(X|G,U)\\ &= \Pr(G)\Pr(Y)\end{aligned}$$

So

$$Y \perp X|U \rightarrow Y \perp G.$$

Didelez & Sheehan. SMMR 2007

Estimating causal effect in linear models

Two more assumptions required for estimation:

- the effect of X on Y is linear,
- no interaction between X and U,

- Suppose the data generating models are

$$\begin{aligned} X &= \alpha_0 + \alpha_1 G + \alpha_2 U + \varepsilon_1, \\ Y &= \theta_0 + \theta_1 X + \theta_2 U + \varepsilon_2. \end{aligned}$$

- we can fit the following reduced models

$$\begin{aligned} E[X|G] &= \alpha_0 + \alpha_1 G, \\ E[Y|G] &= \beta_0 + \beta_1 G, \end{aligned}$$

IV estimators are essentially ratio estimators

- Observe that

$$\begin{aligned} \beta_1 &= E[Y|G = g+1] - E[Y|G = g] \\ &= \theta_1(E[X|g+1] - E[X|g]) + \theta_2(E[U|g+1] - E[U|g]) \\ &= \theta_1\alpha_1. \end{aligned}$$

  Therefore $\theta_1 = \beta_1/\alpha_1$.

- When there is one causal effect, one instrument, the IV estimator can be written as the ratio of two OLS estimator

$$\widehat{\boldsymbol{\beta}}_{IV} = \frac{\hat{\beta}_1}{\hat{\alpha}_1}$$

- The variance of $\hat{\alpha}_1$ is important; highly variable in small samples!

Didelez & Sheehan. SMMR 2007;16:309—330

Instrumental Variable estimation in linear models

This is well developed in Econometrics literature:

- Suppose **G** and **X** have same dimension (both may contain intercept), and confounder U is absorbed in the error ϵ

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

- The usual OLS does not give unbiased estimation for unconfounded effect, because X and ϵ are correlated.

$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\beta + \mathbf{X}^T\epsilon$$

- If the instrument G is independent of error ϵ

$$\mathbf{G}^T\mathbf{Y} = \mathbf{G}^T\mathbf{X}\beta + \mathbf{G}^T\epsilon$$

$$\widehat{\boldsymbol{\beta}}_{IV} = (\mathbf{G}^T\mathbf{X})^{-1}\mathbf{G}^T\mathbf{Y}$$

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}) \sim N\left(0, \sigma^2 Q_{GX}^{-1} Q_{GG} Q_{XG}^{-1}\right)$$

where $Q_{GX} = plim(\mathbf{G}^T\mathbf{X}/n)$, $Q_{GG} = plim(\mathbf{G}^T\mathbf{G}/n)$

This is the same as the ratio estimator in the simple case

Suppose $\mathbf{X} = (1, X)$, $\mathbf{G} = (1, g)$

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{IV} &= (\mathbf{G}^T\mathbf{X})^{-1}\mathbf{G}^T\mathbf{Y} \\ &= (\mathbf{G}^T\mathbf{X})^{-1}(\mathbf{G}^T\mathbf{G})(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{Y} \\ &= \{(\mathbf{G}^T\mathbf{G})^{-1}(\mathbf{G}^T\mathbf{X})\}^{-1}\{(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{Y}\} \end{aligned}$$

It can be verified that

$$\widehat{\boldsymbol{\beta}}_{IV} = \frac{\hat{\beta}_1}{\hat{\alpha}_1}$$

where β_1 is the slope of regressing Y on g, α_1 is the slope of regressing X on g.

Generalized methods of moment

What if $\mathbf{G}$ has more dimension (l) than $\mathbf{X}$ (p)? More equations than the number of parameters.....

$$\bar{g}_n(\boldsymbol{\beta}) = \frac{1}{n}\mathbf{G}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

- If $l == p$, setting $\bar{g}_n(\boldsymbol{\beta}) = 0$ gives methods of moment estimator.
- More generally, for some $l \times l$ matrix $W_n >$0, let

$$J_n(\boldsymbol{\beta}) = n\bar{g}_n(\boldsymbol{\beta})^T W_n \bar{g}_n(\boldsymbol{\beta})$$

- the goal is to set $J_n(\boldsymbol{\beta})$ "close" to zero
-

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{GMM} &= \text{argmin} J_n(\boldsymbol{\beta}) \\ &= \{(\mathbf{X}^T\mathbf{G})W_n(\mathbf{G}^T\mathbf{X})\}^{-1}(\mathbf{X}^T\mathbf{G})W_n(\mathbf{G}^T\mathbf{Y}) \end{aligned}$$

- The scale of W_n does not change $\widehat{\boldsymbol{\beta}}_{GMM}$

What is the optimal W_n?

- Suppose
$$\sqrt{n}\bar{g}_n(\boldsymbol{\beta}) \rightarrow_d \mathcal{N}(0, \Omega)$$
where $\Omega = E(\mathbf{G}_i^T \mathbf{G}_i \sigma^2)$.

- Suppose $W_n \rightarrow_p W_0$, $1/n\mathbf{X}^T\mathbf{G} \rightarrow_p Q$, The asymptotic distribution
$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{GMM} - \boldsymbol{\beta}) \rightarrow_d \mathcal{N}(0, V_{\boldsymbol{\beta}})$$
where $V_{\boldsymbol{\beta}} = (Q^T W_0 Q)^{-1}(Q^T W_0 \Omega W_0 Q)(Q^T W_0 Q)^{-1})$

- In IID cases, the optimal $W_n \rightarrow_p W_0 = \Omega^{-1}$. $W_n = (\frac{1}{n}\mathbf{G}^T\mathbf{G}\hat{\sigma}^2)^{-1}$, and so the optimal estimator is
$$\{(\mathbf{X}^T\mathbf{G})(\mathbf{G}^T\mathbf{G})^{-1}(\mathbf{G}^T\mathbf{X})\}^{-1}(\mathbf{X}^T\mathbf{G})(\mathbf{G}^T\mathbf{G})^{-1}(\mathbf{G}^T\mathbf{Y})$$

Two-stage least squares (2SLS) estimator

- $$\widehat{\beta}_{IV} = \left(\mathbf{X}^T\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{Y}\right),$$
  $$\sqrt{n}(\widehat{\beta}_{IV} - \beta) \sim N\left(0, \sigma^2(Q_{GX}Q_{GG}^{-1}Q_{XG})^{-1}\right)$$
- This is a 2SLS estimator, computationally simple and stable, first compute $\hat{\mathbf{X}}$
  $$\hat{\mathbf{X}} = \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{X}$$
- then regress Y on $\hat{\mathbf{X}}$
  $$\begin{aligned}\widehat{\beta}_{IV} &= (\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T\mathbf{Y}\\ &= [\mathbf{X}^T\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{Y}\end{aligned}$$
- any regression software can be used to get 2SLS estimator, just compute the variance

The fundamental idea observed from 2SLS

Use instrumental variables to extract that variation in intermediate phenotype (exposure) that is independent of confounding variables, and use this part of variation to estimate the causal effect

- The assumptions except the correlation between X and G are not directly testable, because of the presence of unmeasured confounding U.
- There is less appreciation in evaluating and testing these assumptions
- The binary disease outcomes are difficult to work with the concept of IV (Lecture 2).

Caution about assumptions

1. Randomization is approximation at best (untestable)
   - Deviation from "a natural RCT" can be introduced by population stratification, unknown demographic/behavioural/confounders.....

2. Known association between G and X (testable, but genetic associations are weak)
   - Weak genetic instrument and so poor estimation of causal effect, low power

3. No other pathway from G to Y other than through X (exclusion restriction, untestable)
   - Pleiotropy, linkage disequalibrium with other variants that are also related to Y

Relaxed assumptions: adjust for known confounders

Suppose there is a set of known confounders W (population stratification, demographic/behavioral/socioeconomical factor), denote U to be unknown confounders.

1. $G \perp U | W$
2. G correlate with $X | W$
3. $G \perp Y | X, U, W$

- Testing $Y \perp X | W, U$ is equivalent to testing $Y \perp G | W$.
- In linear models, $\theta_1 = \beta_1 / \alpha_1$ still holds

$$
\begin{aligned}
E[Y|X,W,U] &= \theta_1 X + \theta_2 W + \theta_3 U \\
E[X|G,W] &= \alpha_1 G + \alpha_2 W \\
E[Y|G,W] &= \beta_1 G + \beta_2 W
\end{aligned}
$$

all the previous math works!

Overidentifying restrictions and Sargan's test

We can detect pleiotropy and the validity of IV if

- The number of IVs (l) is more than the number of causal effects (p) to be estimated; not all l equations can be exactly zero
- The null hypothesis is $\mathbf{G} \perp (\mathbf{Y} - \mathbf{X}\beta)$
  - instrument is orthogonal to the error term
  - there is no direct effect left once conditional on $\mathbf{X}$
- Sargan's test for 2SLS for l instrumental variables and 1 causal effect

$$\{\mathbf{G}(\mathbf{Y} - \hat{\theta}_{2SLS}\mathbf{X})\}^T\{\hat{\sigma}_2(\mathbf{G})^T\mathbf{G})\}^{-1}\{\mathbf{G}(\mathbf{Y} - \hat{\theta}_{2SLS}\mathbf{X})\} \rightarrow \chi^2(l-1)$$

under the null that all instruments are valid.

Sargan (1958); Small (2007) JASA

J-statistic

- Hansen (1982) gave general results

  $$J_n(\hat{\beta}) = n\bar{g}_n(\hat{\beta})^T \hat{W}_n \bar{g}_n(\hat{\beta}) \to \chi^2(l-p)$$

  as long as $\hat{W}_n$ converges to the optimal W_0 and $\hat{\beta}$ is efficient GMM estimator

- Large J-statistic will reject null hypothesis so that at least one instrument might be invalid.

- report this J-statistic whenever there are overidentifying conditions for IV.

Hansen (1982)

Test the equality of IV estimator and OLS estimator

The null hypothesis is OLS is consistent and fully efficient

- If there is no unmeasured confounding, OLS estimator will be consistent and efficient; IV is consistent under null or alternative
- Large discrepancy between $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{IV}$ suggests that there is confounding and OLS can not be trusted.
- Durbin-Wu-Hausman test

  $$(\hat{\boldsymbol{\beta}}_{IV} - \hat{\boldsymbol{\beta}}_{OLS})^T D^{-1} (\hat{\boldsymbol{\beta}}_{IV} - \hat{\boldsymbol{\beta}}_{OLS}) \rightarrow_d \chi^2(p)$$

  where $D = Var(\hat{\boldsymbol{\beta}}_{IV}) - Var(\hat{\boldsymbol{\beta}}_{OLS})$
- The derivation of the variance comes from the zero correlation between $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{IV} - \hat{\boldsymbol{\beta}}_{OLS}$ under the null.

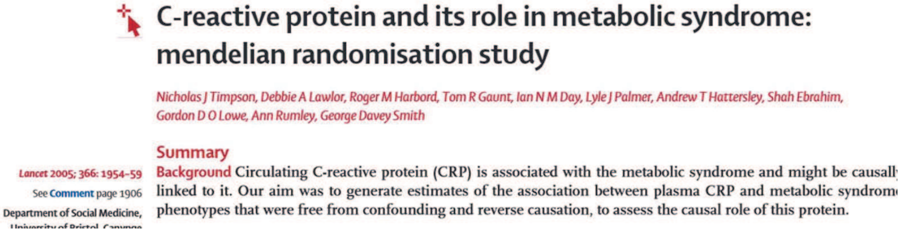
Hausman 1978 Econometrika

Examples of MR

From Jan 2003 to Dec 2013, 179 MR studies were found in PubMed, Medline, Embase and Web of Science (Boef et al (2015) IJE).

- PCSK9 genetic variation related to low LDL cholesterol and decrease coronary heart disease
  - MR analysis suggests causality in Cohen et al 2006 NEJM;354:1264-72
  - Two large RCTs confirmed in 2015 (Sabatine et al NEJM;Robinson et al NEJM)
- Observational studies suggest Lp-PLA$_2$ levels predict CHD
  - MR analysis against causality (Wang et al 2010 Thrombosis Research)
  - Subsequent large RCTs failed to find the benefit (STABILITY investigators NEJM 2014;Nicholls et al 2014 JAMA)
- CRP did not show causal effect on a number of cardiometabolic outcomes; therapies toward CRP are discouraged.
- More examples can be found in Davey Smith 2015 doi: http://dx.doi.org/10.1101/021386.

A MR example



C-reactive protein and its role in metabolic syndrome: mendelian randomisation study

Nicholas J Timpson, Debbie A Lawlor, Roger M Harbord, Tom R Gaunt, Ian N M Day, Lyle J Palmer, Andrew T Hattersley, Shah Ebrahim, Gordon D O Lowe, Ann Rumley, George Davey Smith

Lancet 2005; 366: 1954–59

See Comment page 1906

Department of Social Medicine, University of Bristol, Canynge

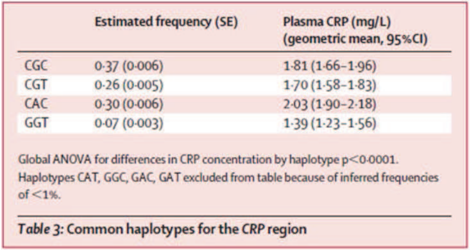
Summary

Background Circulating C-reactive protein (CRP) is associated with the metabolic syndrome and might be causall linked to it. Our aim was to generate estimates of the association between plasma CRP and metabolic syndrom phenotypes that were free from confounding and reverse causation, to assess the causal role of this protein.

Circulating CRP levels are associated with a range of metabolic and cardiovascular diseases (all continuous outcomes in the paper), but not necessarily causal

- CRP haplotype (most likely ones) was used as instrumental variables (likely no other pathway other than circulating CRP)
- CRP haplotype is not associated with potential confounding variables, such as smoking, alcohol, physical activity etc

A MR example



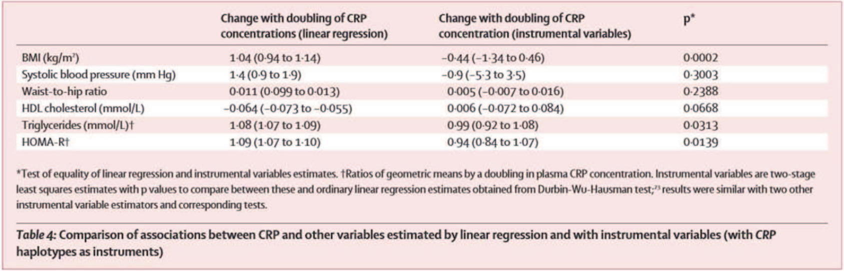
	Estimated frequency (SE)	Plasma CRP (mg/L) (geometric mean, 95%CI)
CGC	0·37 (0·006)	1·81 (1·66–1·96)
CGT	0·26 (0·005)	1·70 (1·58–1·83)
CAC	0·30 (0·006)	2·03 (1·90–2·18)
GGT	0·07 (0·003)	1·39 (1·23–1·56)

Global ANOVA for differences in CRP concentration by haplotype p<0·0001. Haplotypes CAT, GGC, GAC, GAT excluded from table because of inferred frequencies of <1%.

Table 3: Common haplotypes for the *CRP* region

- Strong association between CRP haplotypes and plasma CRP (F-statistic >10), it is not weak instrument
- It would be nice to perform a Sargan's test for validity of instruments.

Difference between MR and observed association



	Change with doubling of CRP concentrations (linear regression)	Change with doubling of CRP concentration (instrumental variables)	p*
BMI (kg/m^2)	1·04 (0·94 to 1·14)	−0·44 (−1·34 to 0·46)	0·0002
Systolic blood pressure (mm Hg)	1·4 (0·9 to 1·9)	−0·9 (−5·3 to 3·5)	0·3003
Waist-to-hip ratio	0·011 (0·099 to 0·013)	0·005 (−0·007 to 0·016)	0·2388
HDL cholesterol (mmol/L)	−0·064 (−0·073 to −0·055)	0·006 (−0·072 to 0·084)	0·0668
Triglycerides (mmol/L)†	1·08 (1·07 to 1·09)	0·99 (0·92 to 1·08)	0·0313
HOMA-R†	1·09 (1·07 to 1·10)	0·94 (0·84 to 1·07)	0·0139

*Test of equality of linear regression and instrumental variables estimates. †Ratios of geometric means by a doubling in plasma CRP concentration. Instrumental variables are two-stage least squares estimates with p values to compare between these and ordinary linear regression estimates obtained from Durbin-Wu-Hausman test;[23] results were similar with two other instrumental variable estimators and corresponding tests.

Table 4: Comparison of associations between CRP and other variables estimated by linear regression and with instrumental variables (with *CRP* haplotypes as instruments)

- IV estimators are computed by 2SLS
- Durbin-Wu-Hausman test for equality of IV and OLS
- These results suggest that there is no causal association between CRP and the metabolic syndrome phenotypes.

Scientific merit of MR studies

Smith & Ebrahim 2003 IJE:

Concluding remark

“For the present, however, it is probably fair to say that the method offers a more robust approach to understanding the effect of some modifiable exposures on health outcomes than does much conventional observational epidemiology. Where possible randomized controlled trials remain the final arbiter of the effects of interventions intended to influence health, however.”

Softwares for IV analysis

- Stata has extensive commands for IV regression: `ivregress`, `ivreg2` implementing 2SLS, Sargan's test or J-statistic, Durbin-Wu-Hausman test

- R package `AER` has `ivreg` function; very powerful `gmm` package

Reference

- Hill AB. The environment and disease: association and causation? *Proceedings of the Royal Society of Medicine*. 1965;58:295-300
- Katan M. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet*.1986;327:507-508.
- Davey Smith G, Ebrahim S. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease. *International Journal of Epidemiology*. 2003;32:1-22.
- Didelez V, Sheehan NA. Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*. 2007;16:309-330.
- Davidson R, MacKinnon J. *Estimation and Inference in Econometrics*. 1993. Oxford University Press, New York.
- Didelez V, Meng S, Sheehan, NA. Assumptions of iv methods for observational epidemiology. *Statistical Science*. 2010;25:22-40.
- Hernan MA, Robins JM. Instruments for causal inference: an epidemiologists dream? *Epidemiology* 2006; 17:360372.
- Hansen, L. 1982. *Large sample properties of generalized method of moments estimators*. Econometrica 50(3): 1029-1054.
- Sargan, J. 1958. *The estimation of economic relationships using instrumental variables*. Econometrica 26(3): 393-415.
- Hausman, J. 1978. *Specification tests in econometrics*. Econometrica 46(3): 1251-1271.
- Small DS. Sensitivity analysis for instrumental variables regression with overidentifying restriction. *JASA* 2007;102:1049-1058.