An Introduction to Mendelian Randomization

Valentin Rousson

Division of Biostatistics
Institute for Social and Preventive Medicine
University Hospital Lausanne
Switzerland

valentin.rousson@chuv.ch

Medical University of Vienna

November 30, 2017

Adiposity vs CRP (C-Reactive Protein, Bochud et al, 2009)

→ COLAUS observational study (Lausanne, 2003-2006, $n = 5362$, 35-75 years)

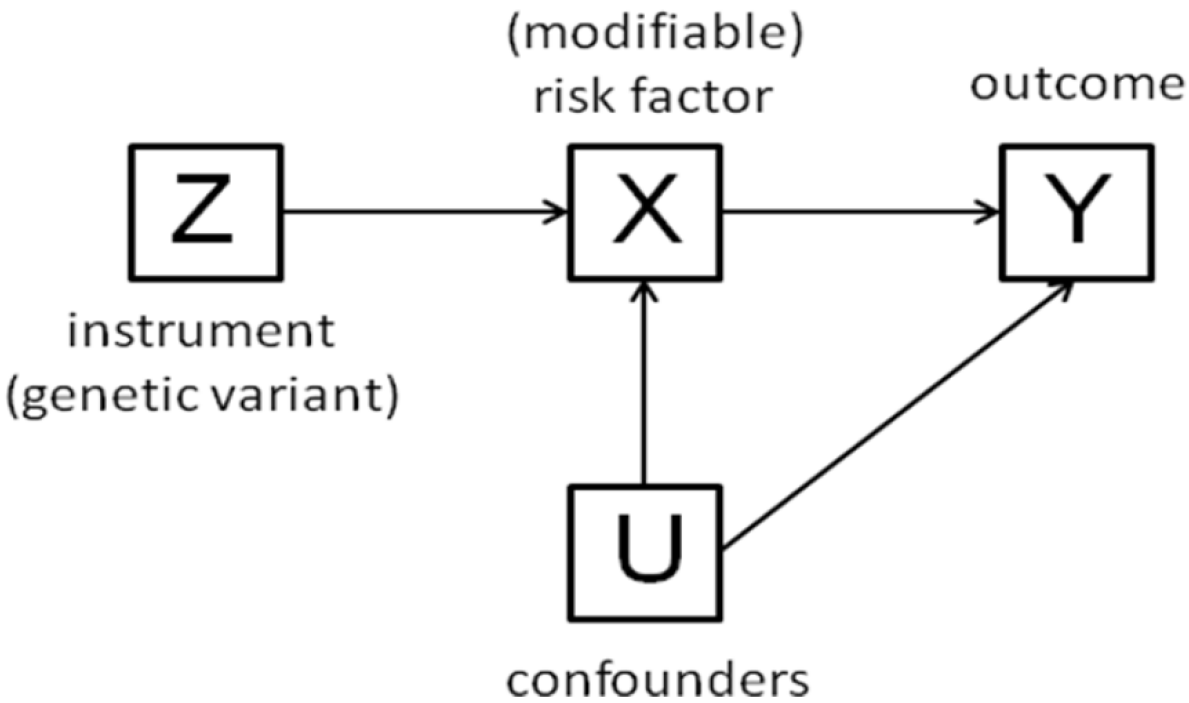
→ adjustment for age, physical activity, regular alcohol consumption, current smoking, hormone replacement therapy (for women)

→ beta regression coefficients (for one-unit increase Log2-CRP, 95% CI) :

response variable	adjusted	men	women
BMI (kg/m^2)	no	0.92 (0.82 ;1.02)	1.44 (1.34 ;1.54)
	yes	0.86 (0.77 ;0.96)	1.35 (1.25 ;1.44)
fat mass (kg)	no	1.78 (1.60 ;1.96)	2.78 (2.59 ;2.97)
	yes	1.50 (1.32 ;1.68)	2.52 (2.33 ;2.70)
lean mass (kg)	no	0.56 (0.36 ;0.77)	0.57 (0.42 ;0.71)
	yes	0.90 (0.69 ;1.11)	0.72 (0.59 ;0.86)

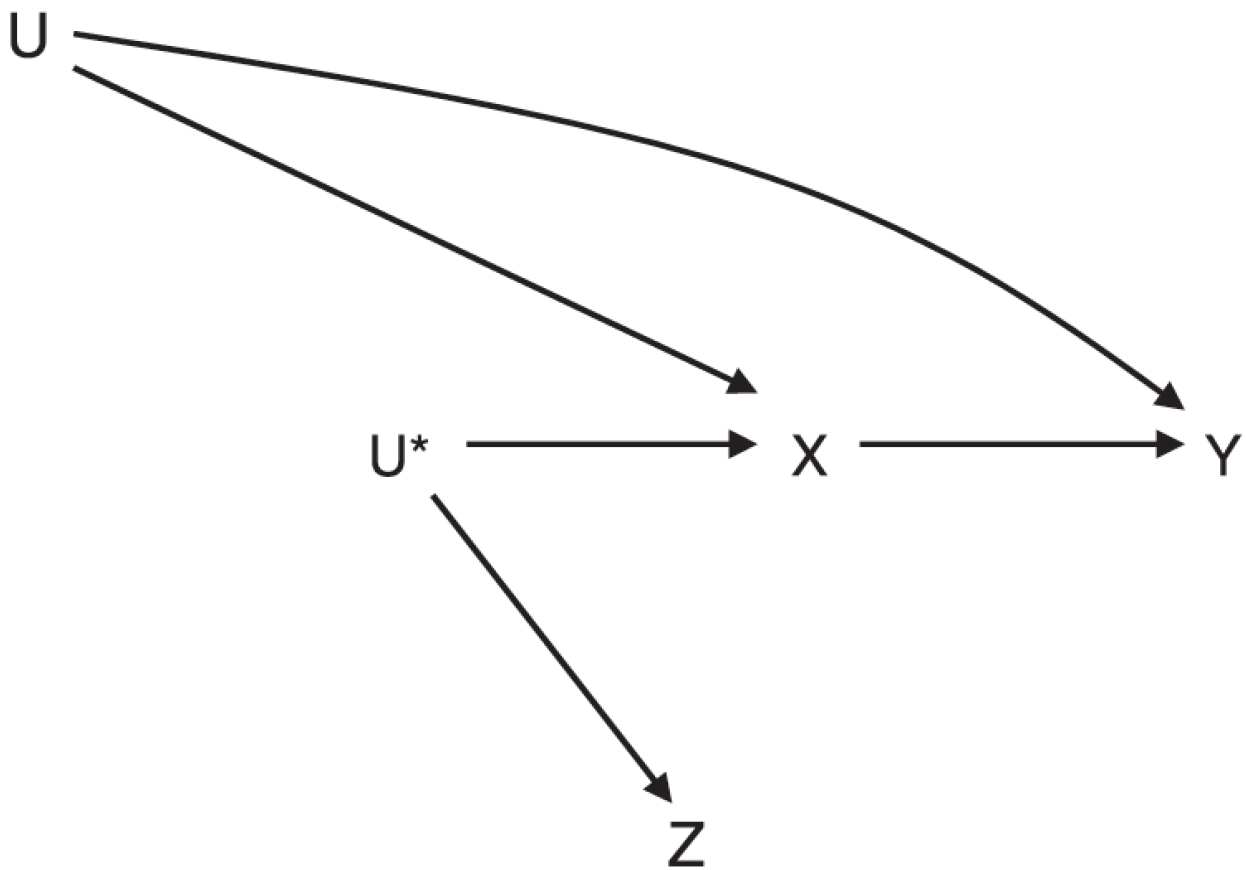
→ although strongly significant associations, not enough to infer a causal effect

Mendelian Randomization (Katan, 1986)



$\rightarrow$ technique of instrumental variables (IV) with genetic information as an instrument

Surrogate instrument (Hernan and Robins, 2006)



$\rightarrow$ U^* is here a surrogate instrument for Z

Method of instrumental variables

$\rightarrow$ core assumptions about the instrument Z :

1. Z is independent from Y given X and U

   $\rightarrow$ no direct pathway from Z to Y

2. Z is independent from any confounder U

   $\rightarrow$ no indirect pathway from Z to Y other than via X

3. Z is (causally) correlated with X

$\rightarrow$ 1+2 : untestable assumptions, judged based on subject matter knowledge

$\rightarrow$ idea : use genetic information Z which is directly (specifically) responsible for X

$\rightarrow$ note : genetic is in principle determined at birth (excluding reverse causation)

$\rightarrow$ under core assumptions : one can test for a causal effect of X on Y by testing for an association between Z and Y (using a classical test)

The Wald method

- (linear) causal relationship between X and Y to be investigated :

$$Y = \alpha + \beta X + U + \varepsilon \quad \cancel{+\gamma Z} \tag{1}$$

$\rightarrow$ one actually has $\gamma = 0$ (assumption 1)

$\rightarrow$ β would be the parameter of interest but cannot be consistently estimated via least-squares with unknown confounders U

- (linear) relationship between Z and X which is assumed :

$$X = \alpha_1 + \beta_1 Z + U_1 + \varepsilon_1 \tag{2}$$

$\rightarrow$ regressing X on Z via least squares is the first step of the method

$\rightarrow$ one obtains an estimate $\widehat{\beta}_1$ of β_1 as Z independent from U_1 (assumption 2)

The Wald method

- (linear) relationship between Z and Y which is assumed :

$$Y = \alpha_2 + \beta_2 Z + U_2 + \varepsilon_2 \tag{3}$$

$\rightarrow$ regressing Y on Z via least squares is the second step of the method

$\rightarrow$ one obtains an estimate $\widehat{\beta}_2$ of β_2 as Z independent from U_2 (assumption 2)

- plugging-in (2) into (1) yields :

$$Y = (\alpha + \beta\alpha_1) + \beta\beta_1 Z + (\beta U_1 + \beta\varepsilon_1 + U + \varepsilon) \tag{4}$$

$\rightarrow$ since Z is independent from U, U_1 and U_2 (assumption 2), one has $\beta_2 = \beta\beta_1$ and thus $\beta = \beta_2/\beta_1$, where $\beta_1 \neq 0$ (assumption 3)

$\rightarrow$ let $\widehat{\beta}_{IV} = \widehat{\beta}_2/\widehat{\beta}_1$ be the third step of the method (consistent estimate of β)

Two-stage least squares

$\rightarrow$ alternatively (equivalently), one can use two-stage least squares

1. fit via least squares :

$$X = \alpha_3 + \beta_3 Z + \varepsilon_3 \tag{5}$$

   $\rightarrow$ let $\widehat{X} = \widehat{\alpha}_3 + \widehat{\beta}_3 Z$ the fitted values of this first model

2. fit via least squares :

$$Y = \alpha_4 + \beta_4 \widehat{X} + \varepsilon_4 \tag{6}$$

   $\rightarrow$ $\widehat{\beta}_4$ $(= \widehat{\beta}_{IV})$ is a consistent IV-estimate of β

$\rightarrow$ one can add further covariates, provided they are in both equations

$\rightarrow$ one can use several instruments (satisfying the core assumptions) in first equation

$\rightarrow$ confidence intervals should take into account uncertainty due to first stage

Adiposity vs CRP (Bochud et al, 2009)

→ genetic instruments : genes rs7553007 and rs1805096

→ 3 possible values (from 0 to 2, number of alleles associated with lower CRP)

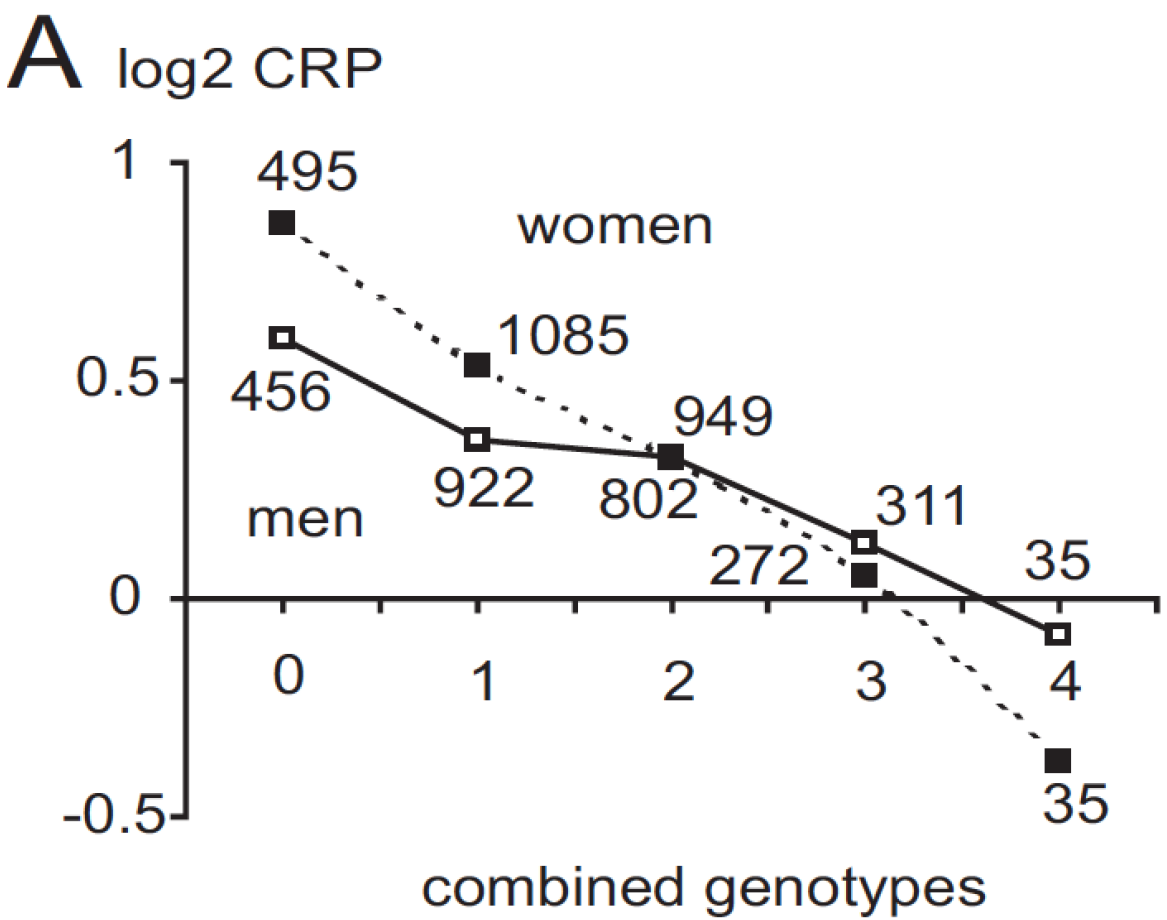
→ genetic score : rs7553007 + rs1805096 (5 possible values from 0 to 4)

→ BMI vs Log2-CRP :

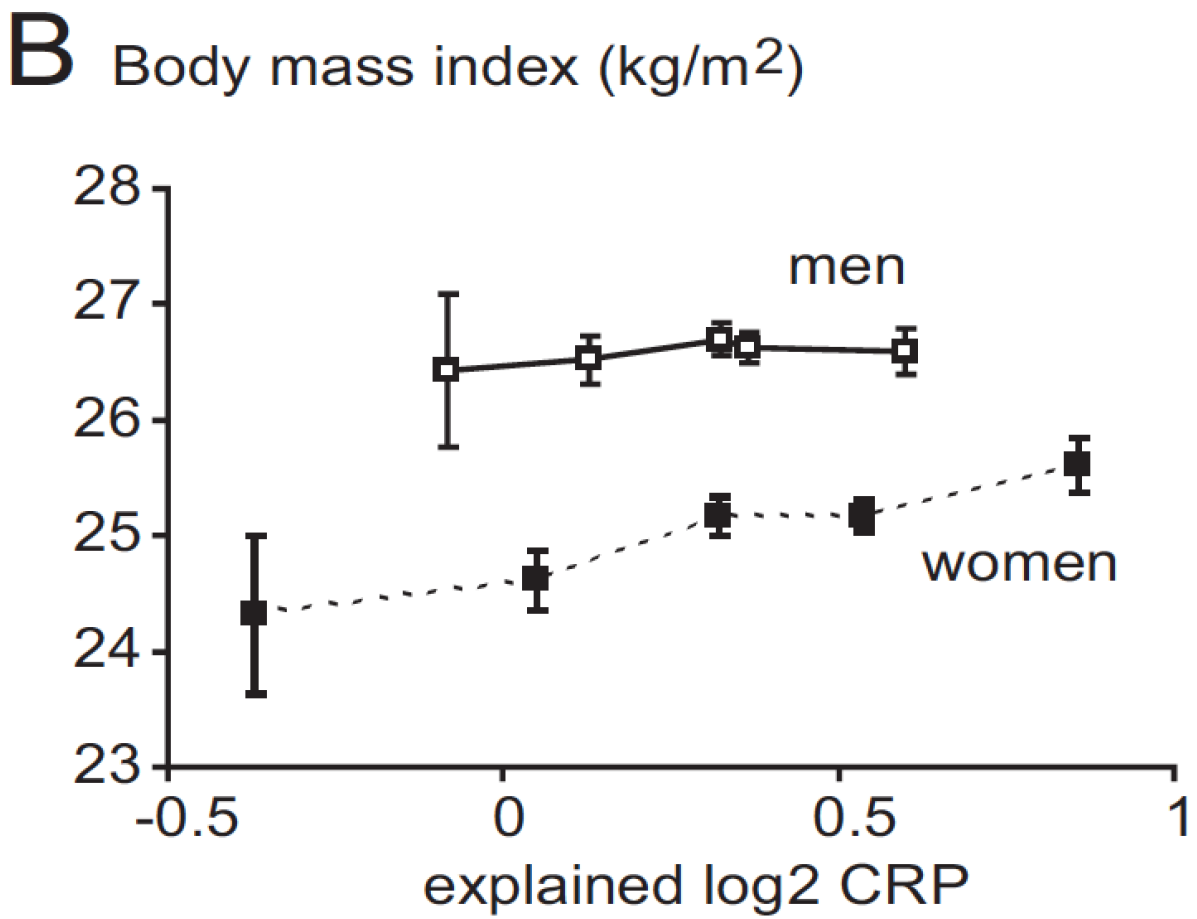
instrument	adjusted	men		women	
none (OLS)	no	0.92 (0.82 ;1.02)	$p < 0.001$	1.44 (1.34 ;1.54)	$p < 0.001$
	yes	0.86 (0.77 ;0.96)	$p < 0.001$	1.35 (1.25 ;1.44)	$p < 0.001$
rs7553007	no	0.12 (−0.79 ;1.02)	$p = 0.80$	1.22 (0.18 ;2.25)	$p = 0.02$
	yes	0.17 (−0.68 ;1.02)	$p = 0.70$	1.29 (0.32 ;2.27)	$p = 0.01$
rs1805096	no	−0.41 (−6.56 ;5.75)	$p = 0.90$	0.78 (−0.12 ;1.67)	$p = 0.09$
	yes	0.16 (−3.60 ;3.92)	$p = 0.93$	0.70 (−0.17 ;1.57)	$p = 0.11$
score	no	0.04 (−1.12 ;1.20)	$p = 0.95$	0.98 (0.32 ;1.63)	$p = 0.004$
	yes	0.14 (−0.90 ;1.18)	$p = 0.79$	0.97 (0.34 ;1.60)	$p = 0.002$

→ suggestion of causal effect of CRP on BMI for women, not for men !

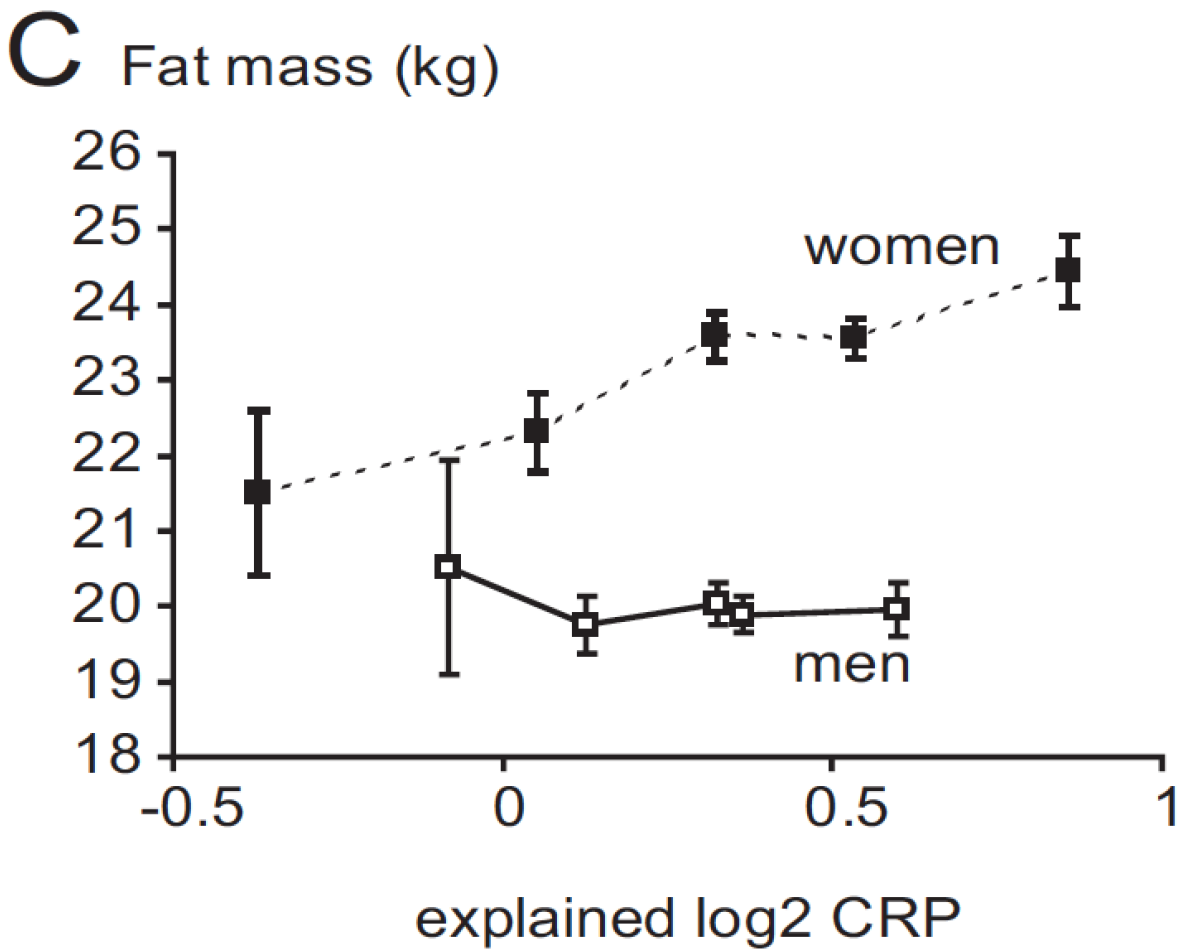
First stage regression



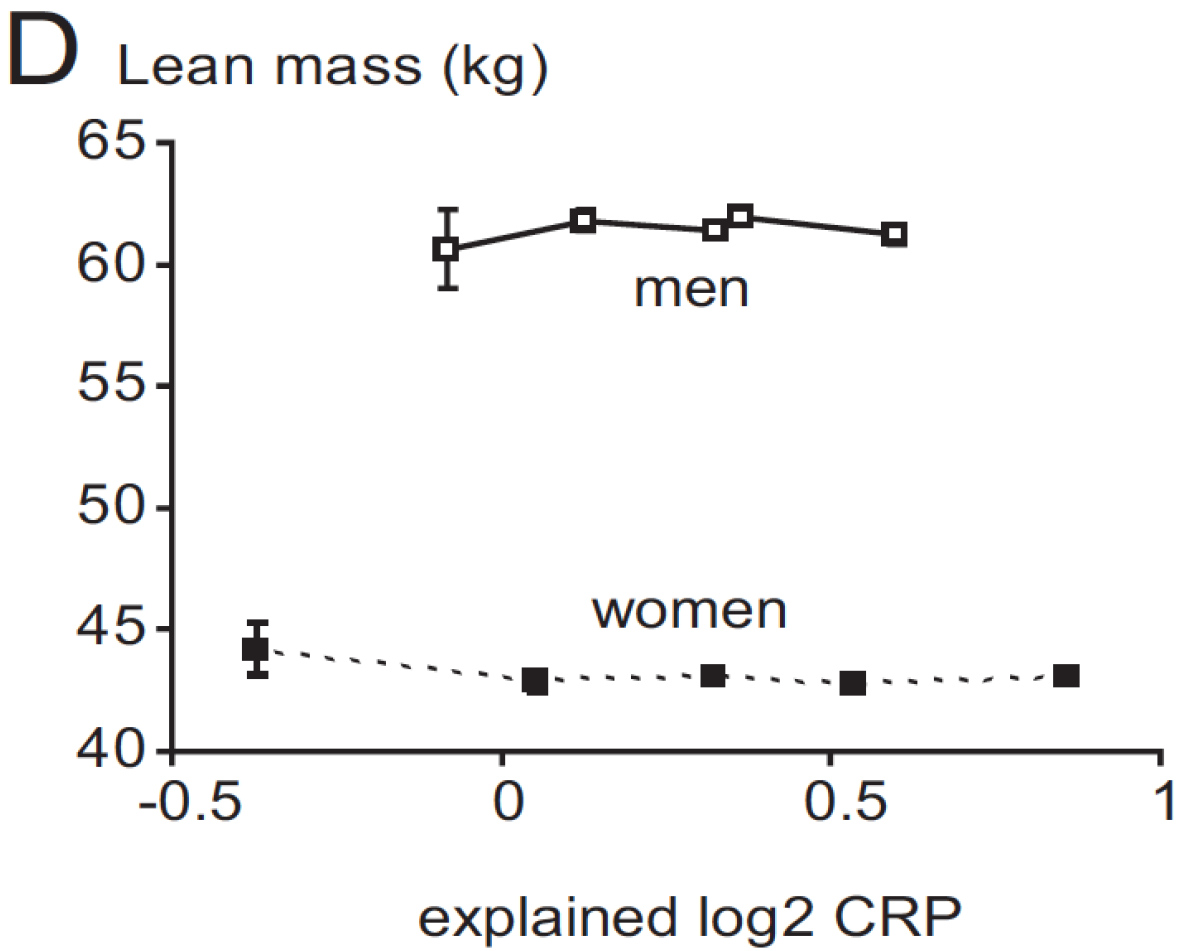
Second stage regression (BMI vs CRP)



Second stage regression (fat mass vs CRP)



Second stage regression (lean mass vs CRP, negative control !)



Checking assumptions 1-2 via DAG (Didelez and Sheehan, 2007)

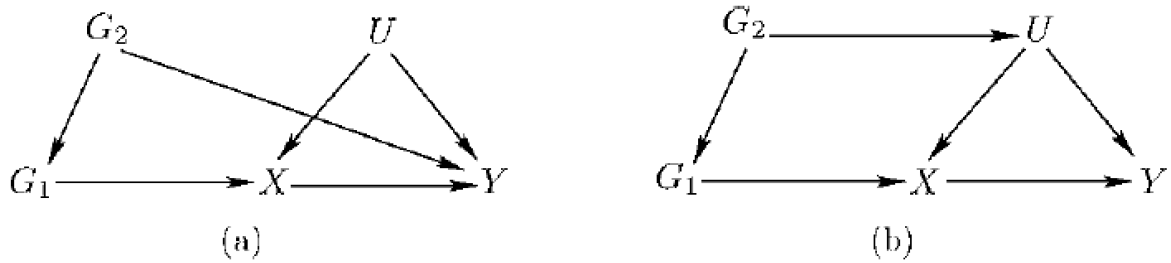


Figure 5 Linkage disequilibrium in a Mendelian randomization application.

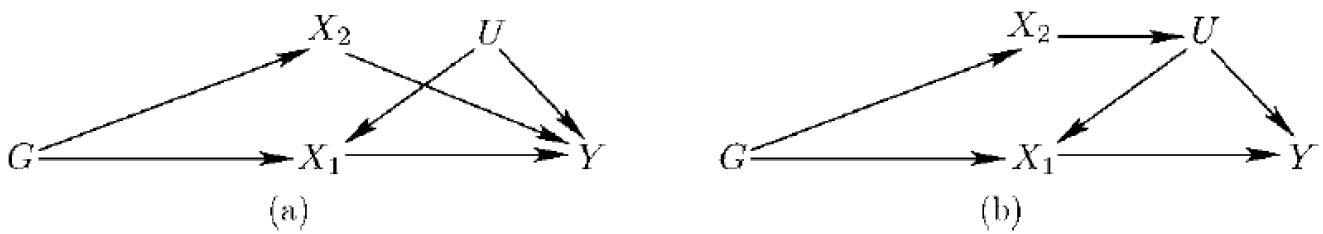


Figure 6 Pleiotropy in a Mendelian randomization application.

Checking assumptions 1-2 via DAG (Didelez and Sheehan, 2007)

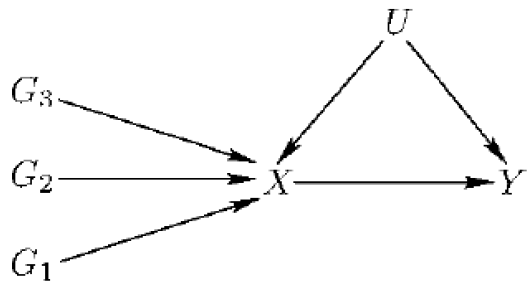


Figure 7 Genetic heterogeneity in a Mendelian randomization application.

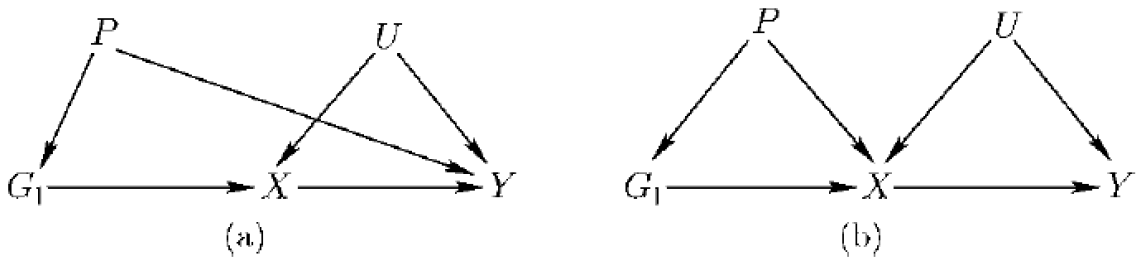


Figure 8 Two examples of population stratification for Mendelian randomization.

Checking (understanding) the method via simulations (Bochud and Rousson, 2010)

$\rightarrow$ all ε_j variables below were simulated as $N(0,1)$

$\rightarrow$ sample size : $n = 100$

$\rightarrow$ five considered situations :

1. (causal effect without confounding, $\beta = 1$)
   $Z = \varepsilon_1$; $X = Z + \varepsilon_2$; $Y = X + \varepsilon_3$

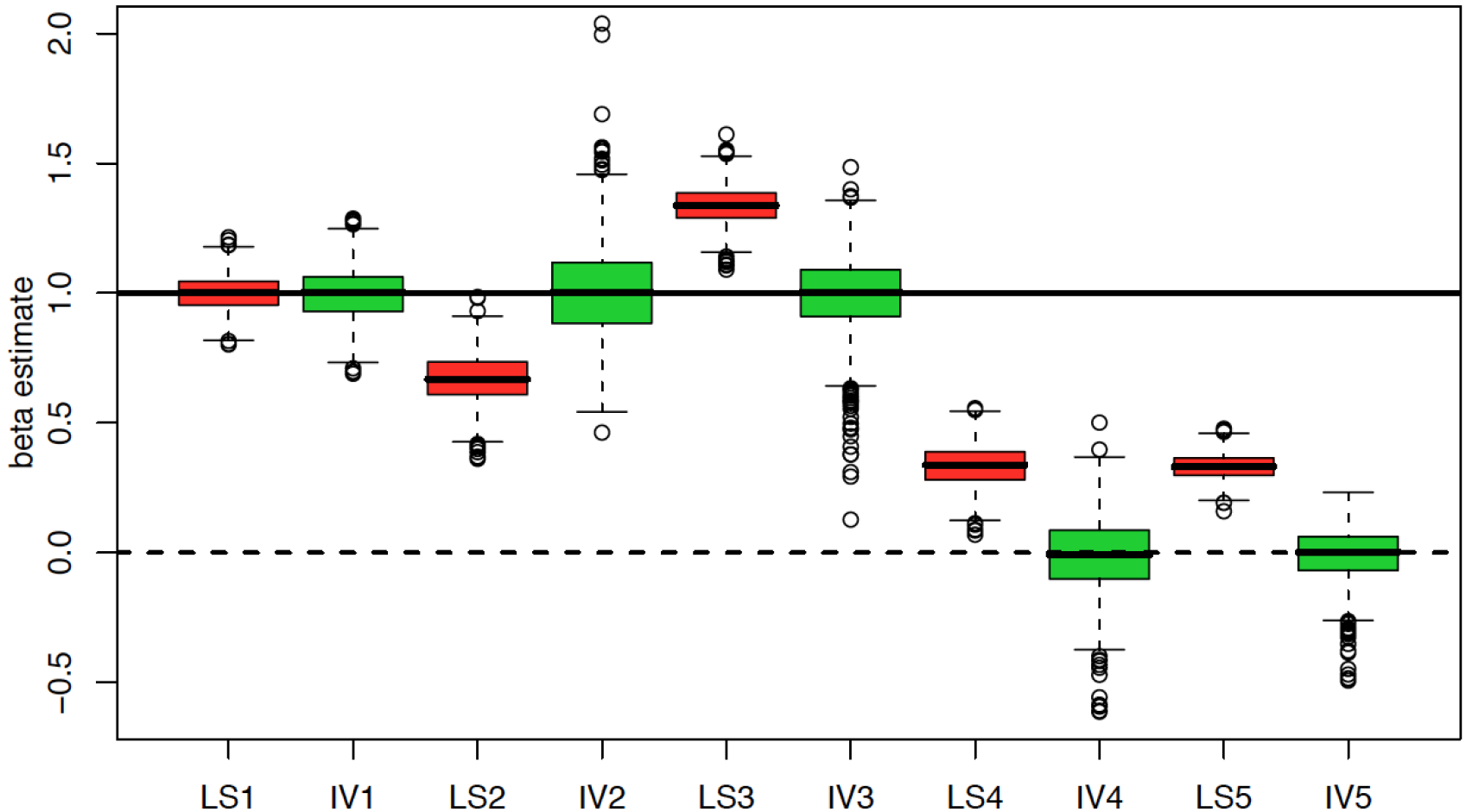
2. (measurement errors, $\beta = 1$)
   $Z = \varepsilon_1$; $X_{true} = Z + \varepsilon_2$; $Y_{true} = X_{true} + \varepsilon_3$; $X = X_{true} + \varepsilon_4$; $Y = Y_{true} + \varepsilon_5$

3. (causal effect with confounding, $\beta = 1$)
   $Z = \varepsilon_1$; $U = \varepsilon_2$; $X = Z + U + \varepsilon_3$; $Y = X + U + \varepsilon_4$

4. (no causal effect with confounding, $\beta = 0$)
   $Z = \varepsilon_1$; $U = \varepsilon_2$; $X = Z + U + \varepsilon_3$; $Y = U + \varepsilon_4$

5. (reverse causation, $\beta = 0$)
   $Z = \varepsilon_1$; $Y = \varepsilon_2$; $X = Z + Y + \varepsilon_3$

Simulation results



Assumption 3 : the problematic of weak instruments

$\rightarrow$ even if it formally holds, one has a weak instrument if small R^2 in first regression

$\rightarrow$ yields the following problems :

- bias of IV-estimate because $E(A/B) \neq E(A)/E(B)$
- huge variability of IV-estimate given by :

$$SE(\widehat{\beta}_{IV}) = SE\left(\frac{\widehat{\beta}_2}{\widehat{\beta}_1}\right) \approx \sqrt{\frac{SE^2(\widehat{\beta}_2)}{\widehat{\beta}_1^2} + \frac{\widehat{\beta}_2^2 SE^2(\widehat{\beta}_1)}{\widehat{\beta}_1^4} - \frac{2\widehat{\beta}_2 Cov(\widehat{\beta}_1, \widehat{\beta}_2)}{\widehat{\beta}_1^3}}$$

  $\rightarrow$ rule of thumb for sample size calculation : divide by R^2 the sample size that would be calculated in an observational study (to reach a given power) !

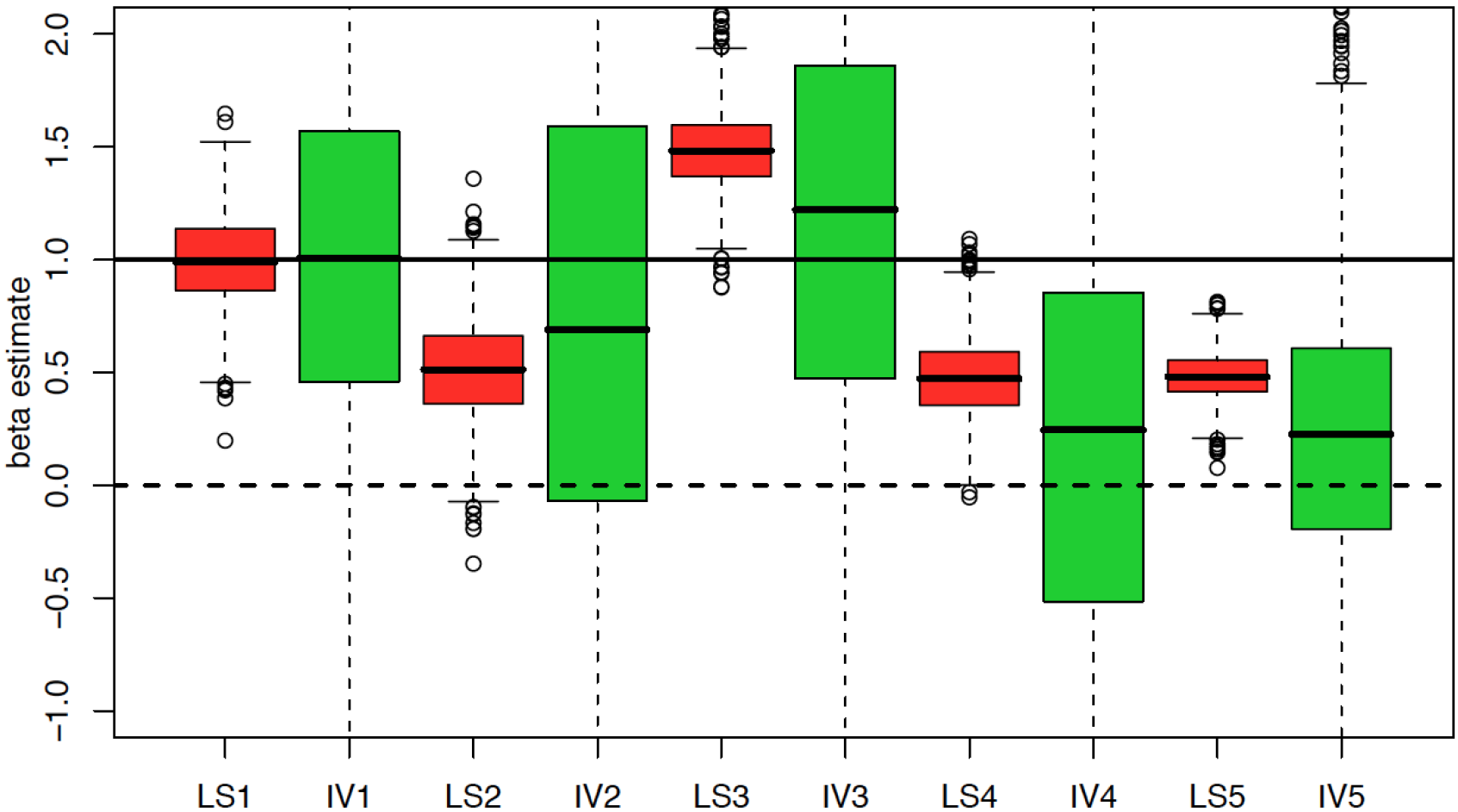
- non-normality (even bimodality !) of IV-estimate (Nelson and Startz, 1990)

$\rightarrow$ rule of thumb to avoid these problems (with p instruments and n individuals) :

$$F = \frac{n-p-1}{p} \cdot \frac{R^2}{1-R^2} > 10$$

(Staiger and Stock, 1997 ; Stock, Wright and Yogo, 2002)

Simulations ($n = 25$, $0.25Z$ instead of Z in design of Bochud and Rousson, 2010)



What if causal effect not constant ?

$\rightarrow$ since already different for men and women, our example suggests that the effect of X on Y might not be constant for all individuals

$\rightarrow$ specifically, this means that one has interactions in model (1)

$\rightarrow$ under that setting, the IV-estimate converges towards a "local average causal effect" (sometimes referred to as LATE in case of a treatment effect)

$\rightarrow$ one needs however an additional "monotonicity assumption" (binary exposure)

$\rightarrow$ in the case of a positive causal relationship between Z and X, this means that for each individual, increasing Z will not decrease X

$\rightarrow$ remember our equation (2) :

$$X = \alpha_1 + \beta_1 Z + U_1 + \varepsilon_1$$

$\rightarrow$ monotonicity assumption will not hold if ε contains measurement errors (or if it is not reproducible, cannot be explained by anything)

$\rightarrow$ partly a philosophical assumption ! (Dawid, 2000)

Binary instrument and binary exposure (Angrist, Imbens and Rubin, 1996)

$\rightarrow$ context of a clinical trial with non-compliance :

- let Z the assigned treatment ($Z = 0$ for placebo, $Z = 1$ for new treatment)
- let X the actual treatment ($X = 0$ for placebo, $X = 1$ for new treatment)

$\rightarrow$ we use the following terminology :

- complier : $X = Z$
- always-taker : $X = 1$ (whatever Z)
- never-taker : $X = 0$ (whatever Z)
- defier : $X \neq Z$

$\rightarrow$ monotonicity assumption : there are no defiers

$\rightarrow$ in that case, IV-estimate converges to average causal effect among compliers

$\rightarrow$ however : compliers is not an identifiable (sometimes small) subset of individuals !

Clinical trial with non-compliance

$\rightarrow$ kind of ideal situation where the untestable assumptions may appear reasonable, and where the instrument should not be weak

$\rightarrow$ let U represent the fact to be complier, always-taker, never-taker or defier :

- let μ_{ij} the mean outcome for individuals with $Z = i$ and $U = j$

  $\rightarrow$ one has here $i = 0, 1$ and $j = C, A, N, D$

  $\rightarrow$ assumption 1 : $\mu_{0j} = \mu_{1j}$ for $j = A, N$

- let $\omega_C, \omega_A, \omega_N$ and ω_D the proportions of compliers, always-takers, never-takers and defiers (such that : $\omega_C + \omega_A + \omega_N + \omega_D = 1$)

  $\rightarrow$ assumption 2 : same proportions for the two groups $Z = 0$ and $Z = 1$

  $\rightarrow$ assumption 3 : $\omega_C > \omega_D$ (or at least $\omega_C \neq \omega_D$)

  $\rightarrow$ strong instrument if ω_C is large, weak instrument of ω_C is low

Clinical trial with non-compliance

→ first regression :

- mean of X in group $Z = 1$: $\omega_C + \omega_A$
- mean of X in group $Z = 0$: $\omega_D + \omega_A$
- mean difference : $\beta_1 = \omega_C - \omega_D$

  → β_1 is the proportion of compliers if there are no defiers

→ second regression :

- mean of Y in group $Z = 1$: $\omega_C\mu_{1C} + \omega_A\mu_{1A} + \omega_N\mu_{1N} + \omega_D\mu_{1D}$
- mean of Y in group $Z = 0$: $\omega_C\mu_{0C} + \omega_A\mu_{0A} + \omega_N\mu_{0N} + \omega_D\mu_{0D}$
- mean difference : $\beta_2 = \omega_C(\mu_{1C} - \mu_{0C}) + \omega_D(\mu_{1D} - \mu_{0D})$

  → β_2 is the parameter which is estimated by the intention-to-treat estimate

→ core assumptions : the IV-estimate converges to :

$$\frac{\beta_2}{\beta_1} = \frac{\omega_C(\mu_{1C} - \mu_{0C}) + \omega_D(\mu_{1D} - \mu_{0D})}{\omega_C - \omega_D}$$

→ monotonicity assumption : without defiers ($\omega_D = 0$), $\beta_2/\beta_1 = \mu_{1C} - \mu_{0C}$

Mendelian randomization with binary outcome (Vuistiner et al, 2012)

→ variables :

- X : alcohol consumption (1=yes, 0=no)
- Y : hypertension (1=yes, 0=no)
- Z : absence of a protective allele in one marker of ALDH2 gene (1=yes, 0=no), supposed to be responsible for a decrease in alcohol consumption

→ (reconstructed) data :

	$Z=0$		$Z=1$	
	$X=0$	$X=1$	$X=0$	$X=1$
$Y=0$	188	444	50	456
$Y=1$	108	284	40	430
U	C or N	A	N	C or A

→ confounded (as-treated) estimate :

$\widehat{\beta}_{AT} = \frac{284+430}{284+430+444+456} - \frac{108+40}{108+40+188+50} = 0.44 - 0.38 = 0.06$ (95% CI : [0.00; 0.11])

Mendelian randomization with binary outcome (Vuistiner et al, 2012)

$\rightarrow$ estimations :

- $\widehat{\omega}_A = \frac{444+284}{444+284+188+108} = 0.71$
- $\widehat{\omega}_N = \frac{50+40}{50+40+456+430} = 0.09$
- $\widehat{\omega}_C = \widehat{\beta}_1 = (1 - 0.09) - 0.71 = 0.20$
- $\widehat{\beta}_{ITT} = \widehat{\beta}_2 = \frac{40+430}{40+430+50+456} - \frac{108+284}{108+284+188+444} = 0.48 - 0.38 = 0.10$
  (95% CI : $[0.06; 0.14]$)
- $\widehat{\beta}_{IV} = \widehat{\beta}_2 / \widehat{\beta}_1 = \frac{0.10}{0.20} = 0.50$ (95% CI : $[0.27; 0.74]$)

$\rightarrow$ some more estimations :

- $\widehat{\mu}_{0A} = \widehat{\mu}_{1A} = \frac{284}{284+444} = 0.39$
- $\widehat{\mu}_{1CA} := \frac{\widehat{\omega}_C \widehat{\mu}_{1C} + \widehat{\omega}_A \widehat{\mu}_{1A}}{\widehat{\omega}_C + \widehat{\omega}_A} = \frac{430}{430+456} = 0.49$
- $\widehat{\mu}_{1C} = \frac{(\widehat{\omega}_C + \widehat{\omega}_A)\widehat{\mu}_{1CA} - \widehat{\omega}_A \widehat{\mu}_{1A}}{\widehat{\omega}_C} = \frac{(0.20+0.71)0.49 - 0.71 \cdot 0.39}{0.20} = 0.83$
- $\widehat{\mu}_{0N} = \widehat{\mu}_{1N} = \frac{40}{40+50} = 0.44$
- $\widehat{\mu}_{0CN} := \frac{\widehat{\omega}_C \widehat{\mu}_{0C} + \widehat{\omega}_N \widehat{\mu}_{0N}}{\widehat{\omega}_C + \widehat{\omega}_N} = \frac{108}{108+188} = 0.36$
- $\widehat{\mu}_{0C} = \frac{(\widehat{\omega}_C + \widehat{\omega}_N)\widehat{\mu}_{0CN} - \widehat{\omega}_N \widehat{\mu}_{0N}}{\widehat{\omega}_C} = \frac{(0.20+0.09)0.36 - 0.09 \cdot 0.44}{0.20} = 0.33$

Causal odds-ratio among compliers (Lui and Chang, 2010)

$\rightarrow$ can be defined and estimated as follows :

$$OR_{IV} = \frac{\mu_{1C}(1-\mu_{0C})}{\mu_{0C}(1-\mu_{1C})} \qquad \widehat{OR}_{IV} = \frac{\widehat{\mu}_{1C}(1-\widehat{\mu}_{0C})}{\widehat{\mu}_{0C}(1-\widehat{\mu}_{1C})}$$

$\rightarrow$ our example :

- IV-estimate :

$$\widehat{OR}_{IV} = \frac{0.83(1-0.33)}{0.33(1-0.83)} = 9.97 \quad (95\%\ \text{CI} : [2.09; 47.42])$$

- ITT-estimate :

$$\widehat{OR}_{ITT} = \frac{0.48(1-0.38)}{0.38(1-0.48)} = 1.50 \quad (95\%\ \text{CI} : [1.25; 1.79])$$

- confounded (as-treated) estimate :

$$\widehat{OR}_{AT} = \frac{0.44(1-0.38)}{0.38(1-0.44)} = 1.28 \quad (95\%\ \text{CI} : [1.02; 1.60])$$

Some conclusions

→ Mendelian randomization : technique of instrumental variables with genetic information as an instrument

→ goal : estimate a causal effect rather than a mere association

→ smart (and desirable) idea rising new problems :

- unverifiable assumptions
- weak instruments (low power, causality less interesting with few compliers)
- extension to binary outcome (odds-ratio) problematic with continuous exposure

→ interesting : these problems largely disappear for a clinical trial with non-compliance

→ kind of paradox : while the original goal of Mendelian randomization was to improve inference in observational studies, trying to get "closer to" clinical trials in this regard, it can be used at the end to further improve inference in clinical trials

→ clinical trials thus remain a gold standard to assess causal inference !

Some references

- Angrist, J.D., Imbens, G.W. and Rubin, D.B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statisical Association*, 91, 444–455.
- Bochud M et al (2009). Association between C-reactive protein and adiposity in women. *Journal of Clinical Endocrinology and Metabolism*, 94, 3969-3977.
- Bochud M, Rousson V (2010). Usefulness of Mendelian Randomization in Observational Epidemiology. *International Journal of Environmental Research and Public Health*, 7, 711-728.
- Burgess S, Small D, Thompson S (2015). A review of instrumental variable estimators for Mendelian randomization. *Satistical Methods in Medical Research*. Ahead of print.
- Dawid A (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95, 407-424.
- Didelez V, Sheehan N (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16, 309-330.
- Hernan M, Robins J (2006). Instruments for causal inference : An epidemiologist's dream ? *Epidemiology* 17, 360-372.
- Katan M (1986). Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet*, 1, 507-508.
- Lawlor D et al. (2008). Mendelian randomization : using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27, 1133-1163.
- Lui K, Chang K (2010). Notes on odds-ratio estimation for a randomized clinical trial with noncompliance and missing outcomes. *Journal of Applied Statistics*, 37, 2057-2071.
- Nelson CR, Startz R (1990). The distribution of the instrumental variables estimator and its ratio when the instrument is a poor one. *Journal of Business*, 63, S125-140.
- Staiger D, Stock J (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65, 557-586.
- Stock JH, Wright JH, Yogo M (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economics Statistics*, 20, 518-529.
- Vuistiner P, Bochud M, Rousson V (2012). A comparison of three methods of Mendelian randomization when the genetoc instrument, the risk factor and the outcome are all binary. *PLoS ONE*, vol. 7, no. 5, e35951.