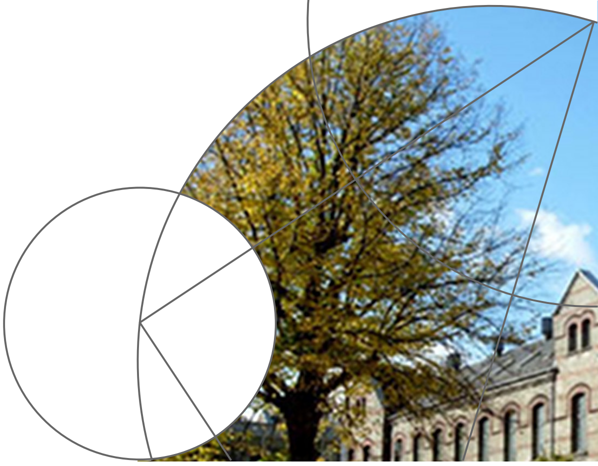
UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Faculty of Health Sciences

Instrumental variables analysis, still room for development? DES-meeting Apr. 2018

Torben Martinussen
Department of Biostatistics

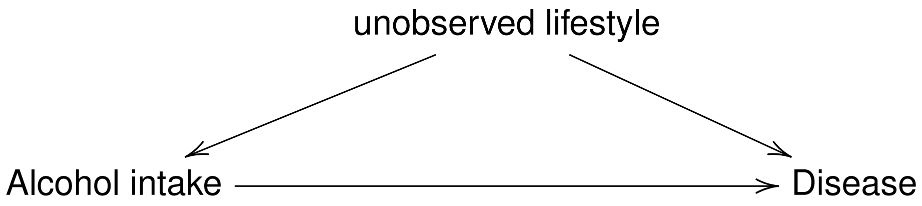


Motivation

- Can you guarantee that the results from your observational study are unaffected by unmeasured confounding?
- The only answer an epidemiologist can provide is "no".
- Imagine for a moment the existence of an alternative method that allows one to make causal inference from observational studies even if the confounders remain unmeasured.
- That method would be an epidemiologist's dream.
- The above quotes are taken from Hernan and Robins (Epidemiology, 2006).

Alcohol consumption

- Can we estimate the exposure effect when there is un-observed confounding?



- Alcohol consumption has been found in observational studies to have positive effects (coronary heart disease) as well as negative effects (liver cirrhosis, some cancers).
- But also strongly associated with all kinds of confounders (lifestyle etc.) as well as subject to self-report bias. Hence doubts in causal meaning of above effects.

Mendelian Randomization: Basic idea

Idea: if we cannot randomise, let's look for instance where nature has randomised, eg though genetic variation.

Example: Alcohol consumption

Genotype: ALDH2 determines blood acetaldehyde, the principal metabolite for alcohol.

Two variants, 1 and 2.

22 homozygotes individuals suffer facial flushing, nausea, drowsiness and headache after alcohol consumption

They hence have low alcohol consumption regardless of other lifestyle behaviours.

IV-idea: check if these individuals have a different risk than others for alcohol related health problems.

Association between alcohol intake and ALDH2 genotype.

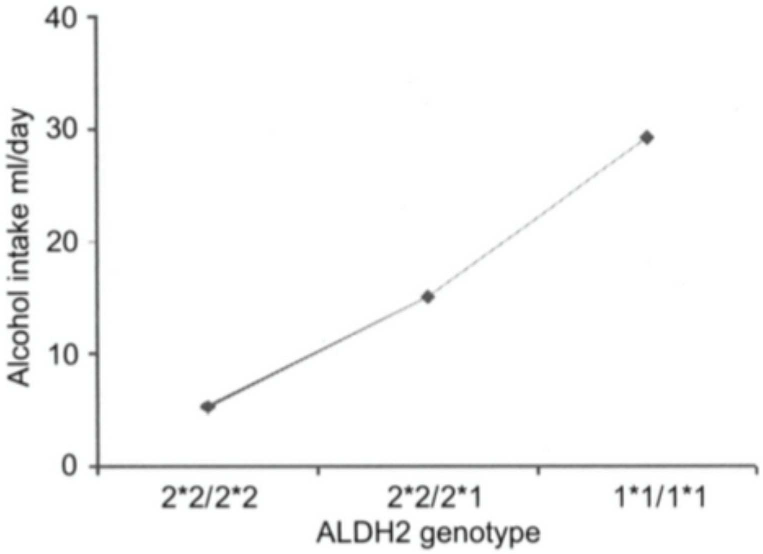
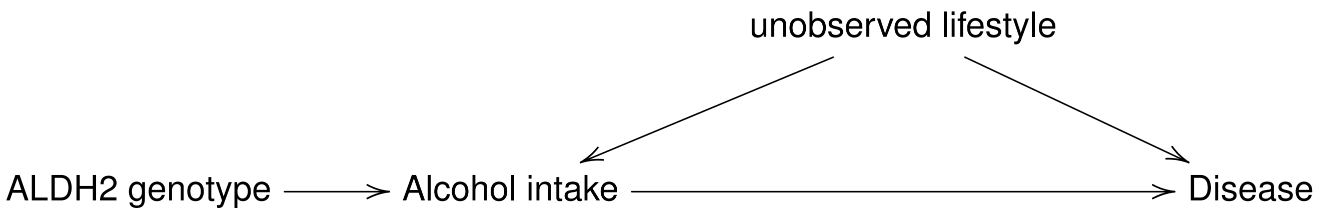


Fig. 1. Relationship between alcohol intake and ALDH2 genotype. Data from Takagi et al. (2002).

Although we see an association here, age and cigarette smoking are not are not related to the genotype! They are of course related to alcohol consumption, however.

Alcohol consumption



unobserved lifestyle

ALDH2 genotype

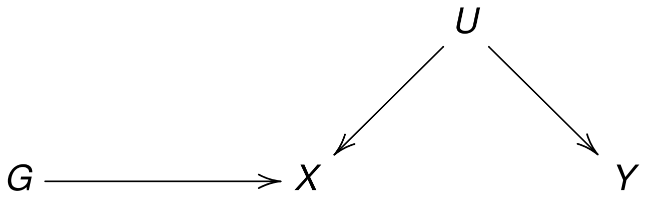
Alcohol intake

Disease

- Causal effect? Under IV-assumptions, the null-hyp of no causal effect of alcohol consumption, should imply no association between ALDH3 and disease.
- While if alcohol consumption has a causal effect we would expect an association between ALDH2 and disease.

Using IV to test for causal effect

- Simply test if Y (outcome) and G (instrument) are independent.



- Regardless of measurement level, testing independence between Y (outcome) and G (instrument) is valid test for presence of causal effect of X (exposure) on Y.

Alcohol consumption

Results (Meta-analysis by Chen et al. (2008).

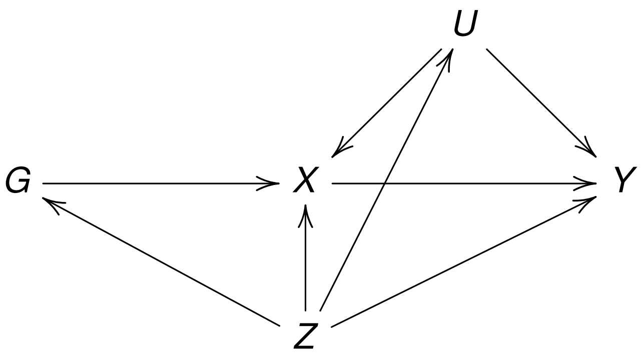
- Blood pressure on average 7.44 mmHg higher and risk of hypertension 2.5 higher for ALDH2 11's than 22's carriers (only males).
  - Mimics the the effect of large versus low alcohol consumption.
- Blood pressure on average 4.24 mmHg higher and risk of hypertension 1.7 higher for ALDH2 12's than 22's carriers (only males).
  - Mimics the the effect of moderate versus low alcohol consumption.
  - indicates that even moderate alcohol consumption is harmful.

Mendelian Randomization

- This was an example of a Mendelian Randomization analysis
- **It is not about** establishing a causal relationship between gene and disease.
- The goal is to investigate whether there is a **causal relationship** between a **modifiable risk factor** (alcohol consumption) and the **disease**.
- Mendelian Randomization analysis is an example of what we more generally call instrumental variables analysis.

Instrumental variables DAG

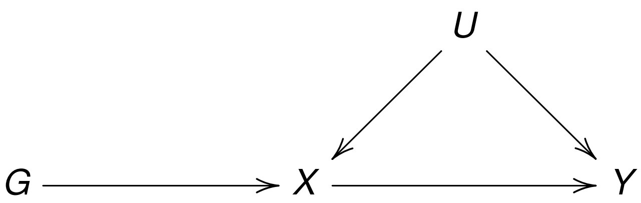
- Can we estimate the exposure effect when there is un-observed confounding?



Figur: G is is the instrument, X the exposure variable. U: potential unmeasured confounders; observed confounders by Z.

Instrumental variables DAG

Let's look at DAG without additional observed covariates Z:



Core assumptions:

1. G and X are associated
2. G is independent of the unmeasured confounder(s) U
3. G is independent of the outcome given X and U. (Exclusion restriction).

Ass. 2. and 3. are hard (often impossible) to check from the observed data.

OBS: The 3 ass. above are actually not enough to do estimation!

Examples of instruments

- Genes; Mendelian randomization.
- Treatment assignment in randomized clinical trials with non-compliance.
- Physician treatment prescribing preference.
- Martens et al. (Epi, 2006).

Using IV to estimate causal parameters

- Target parameters (continuous response; binary response; time-to-event response).
- Classical 2-stage-least square estimator.
- Methods for binary outcome data
- Methods for time-to-event outcome data
- Methods for competing risk data.

Continuous outcome Y

We will assume the (underlying true) model

$$E(Y|X, G, Z, U) = \beta_0 + \beta_X X + \beta_Z Z + \beta_U U$$

If we had information about the confounder(s) U we could just do ordinary regression

```
> n=10000
> U=rnorm(n)
> G=rbinom(n,1,0.5)
> Z=rbinom(n,1,0.5)
> alp0=1;alp1=0.75;alp2=-0.5;alp3=0.3
> mu=alp0+alp1*G+alp2*U+alp3*Z
> X=rnorm(n,mu,1)
>
> bet0=1;bet1=0.5;bet2=-1;bet3=0.3
> thet=bet0+bet1*X+bet2*U+bet3*Z
> Y=rnorm(n,thet,0.5)
```

Hence, the β_X is 0.5 in this scenario. Can we estimate it from data?

Continuous outcome Y

If we had information about the confounder(s) U we could just do ordinary regression:

```
> summary(lm(Y~X+Z+U))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.988966   0.009615  102.86   <2e-16 ***
X            0.503275   0.004709  106.88   <2e-16 ***
Z            0.321210   0.010139   31.68   <2e-16 ***
U           -0.997307   0.005549 -179.72   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Problem is that we don’t get to see U!

So we cannot just do the above regression.

Continuous outcome Y

Do the regression without U, that is, use the data we actually get to see (so this may be the problem with an observational study):

```
> summary(lm(Y~X+G+Z))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.575063   0.019928   28.86   <2e-16 ***
X            0.894512   0.009174   97.51   <2e-16 ***
G           -0.280218   0.021542  -13.01   <2e-16 ***
Z            0.220605   0.020671   10.67   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Problem is that we now get a biased estimate of effect of X; remember true coefficient was 0.5.

Note also that we see a significant effect of G in the above analysis. But we know that there should be no such one. Why this contradictory result?

Continuous outcome Y

Is there a causal effect of X?

```
> summary(lm(Y~G))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.72321    0.02038   84.56   <2e-16 ***
G            0.37182    0.02903   12.81   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

- Can we somehow estimate the correct effect of X without using U (which we don’t observe)?
- Yes, we can! Under the core assumptions (and actually a little more).
- It is called 2SLS estimation
- In this scenario the so-called Wald estimator also gives the same.

Continuous outcome Y

So what are these estimates?

The Wald estimator:

- Do the regression of Y on G (and Z), $\hat{\beta}_{Y|G}$
- Do the regression of X on G (and Z), $\hat{\beta}_{X|G}$
- The Wald estimator is then defined as (ITT-estimator inflated)

$$\frac{\hat{\beta}_{Y|G}}{\hat{\beta}_{X|G}} \left(= \frac{E(Y|G=1) - E(Y|G=0)}{E(X|G=1) - E(X|G=0)} \right)$$

```
> lm(Y~G+Z)
Coefficients:
(Intercept)            G            Z
     1.4770       0.3700       0.4924
> lm(X~G+Z)
Coefficients:
(Intercept)            G            Z
     1.0083       0.7269       0.3039
> wald=coef(lm(Y~G+Z))[2]/coef(lm(X~G+Z))[2]
> wald
        G
0.5090283
```

Continuous outcome Y

So what are these estimates?

The 2SLS estimator:

- Do the regression of X on G (and Z), and calculated the predicted values: $\hat{X}$
- Do the regression of Y on $\hat{X}$ (and Z)

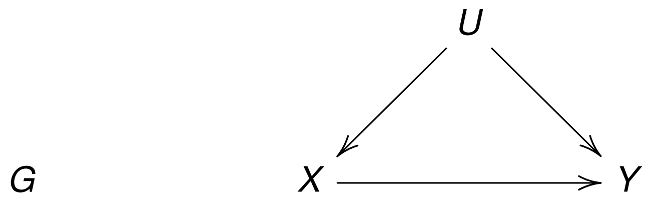
```
> hat.X=fitted(lm(X~G+Z))
> summary(lm(Y~hat.X+Z))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.96376    0.05743   16.78   <2e-16 ***
hat.X        0.50903    0.03936   12.93   <2e-16 ***
Z            0.33775    0.03105   10.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So again, we get the estimate 0.509.

The above reported s.e. is incorrect, but a correct one is available.

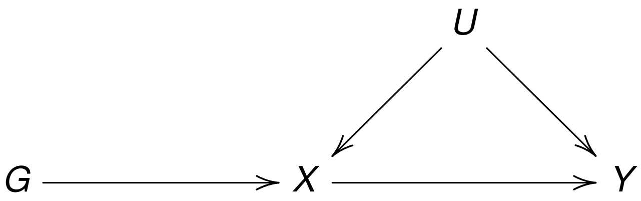
Strength of instrument

- It is crucial that there is some correlation between instrument and exposure
- If they are un-correlated there is no hope of estimating the causal effect of the exposure.



- Any bias will be blown up.
- An F-statistic below 10 is considered a situation where the instrument is weak.
- Paper by Jackson and Swanson (Epi, 2016, Rothman prize winner). $corr(G, U)$ should be scaled with $1/corr(X, G)$ using as U all the observed confounders, before comparing with $corr(X, U)$.
- See also Davies et al. (IJE, 2017) on a follow up on the bias-plot procedure of Jackson and Swanson. Remember to take sampling variability into account.

Let’s have a look at the core assumptions



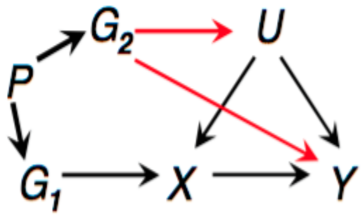
Core assumptions:

1. G and X are associated
2. G is independent of the unmeasured confounder(s) U
3. G is independent of the outcome given X and U.

Number 2. and 3. are untestable.

Linkage disequilibrium

Graphical representation: G_1 = chosen instrument, in LD with other gene G_2 through parental genes P.

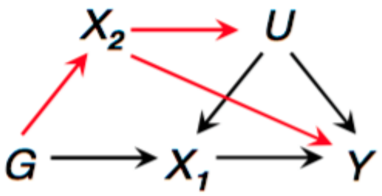


Here $Y \not\perp\!\!\!\perp G|(X,U)$ or $G \not\perp\!\!\!\perp U$!

Pleiotropy

Pleiotropy refers to a genetic variant having multiple function. The chosen gene might not only affect the phenotype of interest but also other traits.

Graphical representation:
G affects another phenotype X_2 (phenotype of interest is X_1)



Here, again, $Y \not\perp\!\!\!\perp G | (X, U)$ or $G \not\perp\!\!\!\perp U$.

Binary response data

- We have now seen an IV-method for continuous outcome, the 2SLS approach
- Does something similar exist for for binary outcome data?
- Could you imagine a 2SLS approach for binary outcome data?
- How would you do it?

Binary response data

- Yes, so surely it is possible to carry out 2SLS also when the outcome is binary
- Unfortunately, it does not estimate any sensible causal parameter.
- Unless we make linear model for $p = P(Y = 1|X, U)$ which is non-standard.
- Usually we wish to use logistic regression when having a binary response (reporting OR's).
- Still, one can carry out a 2SLS method, but, as mentioned, it does not estimate any sensible causal paramter.
- See eg Vansteelandt et al. (Stat Sci, 2011)

Binary response data

- The OR is target parameter.
- There is a suggestion by Vansteelandt and Goetghebeur (2003), the double-logistic approach.
- Target parameter is β in

$$\text{logit}\{P(Y^x = 1|X = x, G)\} - \text{logit}\{P(Y^0 = 1|X = x, G)\} = \beta x$$

- Causal effect among the treated ($x = 1$); causal effect among the exposed.

Time-to-event data

- Assume the outcome is a (possibly right-censored) time-to-event outcome
- Example: time-to-death; exposure vitamin D.
- Not so much work concerning IV analysis for this type of outcome has been done.
- This has been within my own research area.
- Papers: Tchetgen Tchetgen et al. (Epidemiology, 2015); Martinussen et al. (Biometrics, 2017).
- Both use additive hazards models (absolute risk).
- In the first one we develop a 2SLS method, but it is based on stronger assumptions than the last one.
- In Martinussen et al. (Biostatistics, 2017), we develop a method using an IV in Cox-regression setup.

Unmeasured confounding and competing risks

- Time-to-event analyses are often plagued by both – possibly **unmeasured – confounding** and **competing risks**.

- HIP trial on effectiveness of screening on breast cancer mortality:
  - About 60000 women aged 40-60 were randomized into two equally sized groups.
  - Approximately 35% of the women in the screening group refused to participate (non-compliers) There were large differences between the study women who participated and those who refused (Shapiro, 1977)
  - Results from the "as treated"analysis may be doubtful due to unobserved confounding.
  - There is further a competing risk issue in these data. In the first 10 years of follow-up there are 4221 deaths but only 340 were deemed due to breast cancer.
  - An IV-analysis is tempting as the original randomization can be used as instrument.
  - How can we do this when dealing with competing risk data?

Competing risks and IV

Let $(\tilde{T}^x, \delta^x)$ denote the counterfactual event time and event type that would be observed for given subject if the exposure of that subject was set to x.

Target of inference is thecontrast between the counterfactual cause-specific hazard functions:

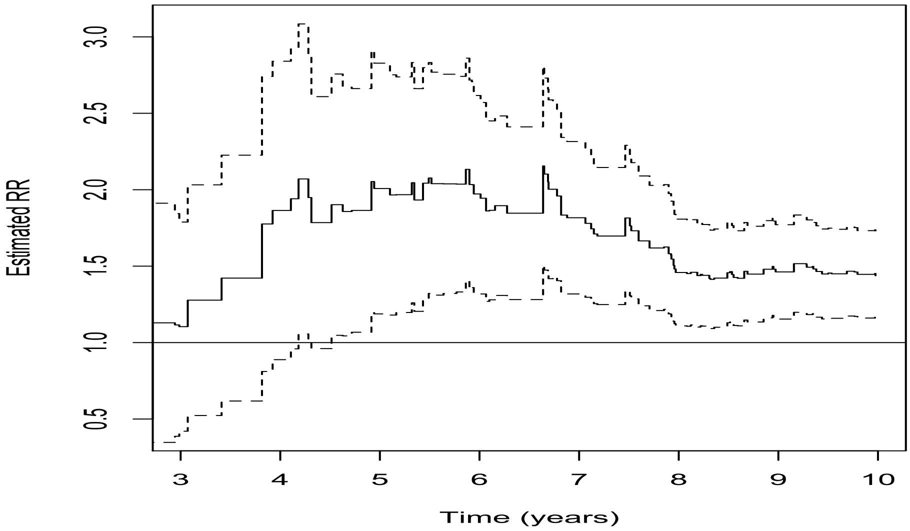
$$\lambda^j_{T^x}(t|X = x, G, L) - \lambda^j_{T^0}(t|X = x, G, L) = \beta_j(t)x,$$

for $j = 1, 2$

Competing risks and IV

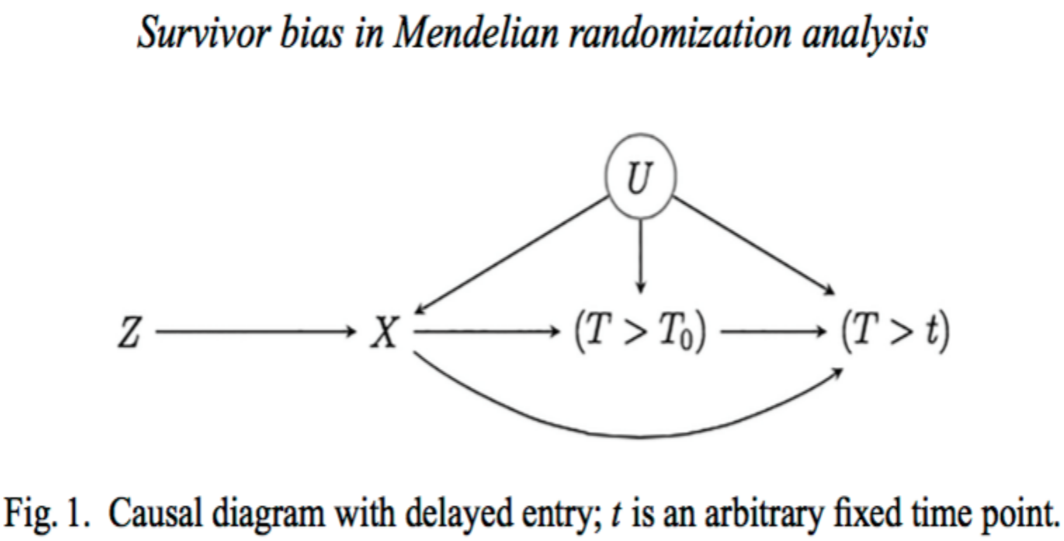
Using our method we may then estimate

$$\mathrm{RR}(t) \equiv \frac{P(T^0 \leq t, \delta^0 = 1 | X = 1)}{P(T^1 \leq t, \delta^1 = 1 | X = 1)}$$



Survivor bias

Vansteelandt et al. (2017, Biostatistics):



Survivor bias in Mendelian randomization analysis

Fig. 1. Causal diagram with delayed entry; *t* is an arbitrary fixed time point.

Concluding remarks

- Untestable assumption (no unmeasured confounders) is replaced with other un-testable assumptions.
- The fourth assumption. Which is the most reasonable one? Hernan and Robins (Epi, 2006).
- Bias-plot procedure (Jackson and Swanson, Epi, 2016), Davies et al. (IJE, 2017)
- Active research area, MR GENIUS (Tchetgen Tchetgen et. al, Arxiv, 2016). Avoids the exclusion restriction assumption, leads to a new type of 2SLS-estimator.
- MR GENIUS : Mend. Rand. G-Estimation under No Interaction with Unmeasured Selection.
- IV-methods for binary outcome data, and for time-to-event data.
- Competing risk data. Martinussen and Vansteelandt (Arxiv, 2017).
- More practical experience with new these methods is needed.