Some topics on causal inference

Krista Fischer

Estonian Genome Center, University of Tartu, Estonia

Statistical Practice in Epidemiology, Lyon 2018

Statistical associations vs causal effects in epidemiology

Does the exposure (smoking level, obesity, etc) have a causal effect on the outcome (cancer diagnosis, mortality, etc)?

is not the same question as

Is the exposure associated with the outcome?

Conventional statistical analysis will answer the second one, but not necessarily the first.

What is a causal effect?

There is more than just one way to define it.

A causal effect may be defined:

- At the individual level:
  Would my cancer risk be different if I were a (non-)smoker?
- At the population level:
  Would the population cancer incidence be different if the prevalence of smoking were different?
- At the *exposed subpopulation level*:
  Would the cancer incidence in smokers be different if they were nonsmokers?

None of these questions is “mathematical” enough to provide a mathematically correct definition of causal effect

Causal effects and counterfactuals

- Defining the causal effect of an observed exposure always involves some counterfactual (what-if) thinking.
- The individual causal effect can be defined as the difference

$$Y(X = 1) - Y(X = 0)$$

. where $Y(1) = Y(X = 1)$ and $Y(0) = Y(X = 0)$ are defined as individual's potential (counterfactual) outcomes if this individual's exposure level X were set to 1 or 0, respectively.

- Sometimes people (e.g J. Pearl) use the "do" notation to distinguish counterfactual variables from the observed ones: $Y(do(X = 1))$ and $Y(do(X = 0))$.

The "naïve" association analysis

- With a binary exposure X, compare average outcomes in exposed and unexposed populations:

$$E(Y|X=1) - E(Y|X=0)$$

Is cancer incidence different in smokers and nonsmokers?

- But mostly:

$$E(Y|X=1) \neq E(Y(1))$$

Cancer risk in smokers is not the same as the potential cancer risk in the population if everyone were smoking

- Similarly:

$$E(Y|X=0) \neq E(Y(0))$$

- In most cases there is always some unobserved confounding present and therefore the naïve analysis does not provide causal effect estimates.

Counterfactual outcomes in different settings

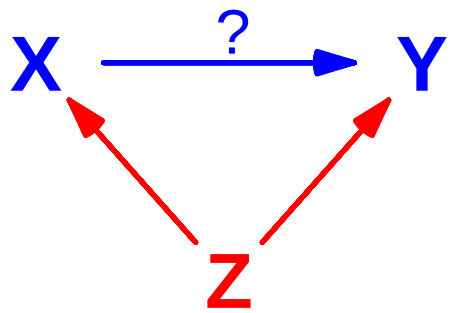
- Randomized trials: probably the easiest setting to imagine $Y(X)$ for different X
- "Actionable" exposures: smoking level, vegetable consumption, . . . – potential interventions may alter exposure levels in future.
- Non-actionable exposures: e.g genotypes. It is difficult to ask *"What if I had different genes?"*. Still useful concept to formalize genetic effects (heritability, attributable risk).
- Combinations: With X– a behavioral intervention level, Z–smoking level and Y–a disease outcome, one could formalize the effect of intervention on outcome by using $Y(X, Z(X))$

Classical/generalized regression estimates vs causal effects?

- In the presence of confounding, regression analysis provides a biased estimate for the true causal effect
- To reduce such bias, one needs to collect data on most important confounders and adjust for them
- However, too much adjustment may actually introduce more biases
- Causal graphs (Directed Acyclic Graphs, DAGs) may be extremly helpful in identifying the optimal set of adjustment variables

Adjustment for confounders I

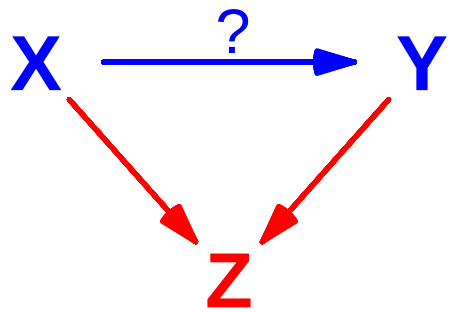
"Classical" confounding: situation where third factors Z influence both, X and Y



For instance, one can assume: $X = Z + U$ and $Y = Z + V$, where U and V are independent of Z.
X and Y are independent, conditional on Z, but marginally dependent.
One should adjust the analysis for Z, by fitting a regression model for Y with covariates X and Z. There is a causal effect between X and Y, if the effect of X is present in such model.

Adjustment may sometimes make things worse

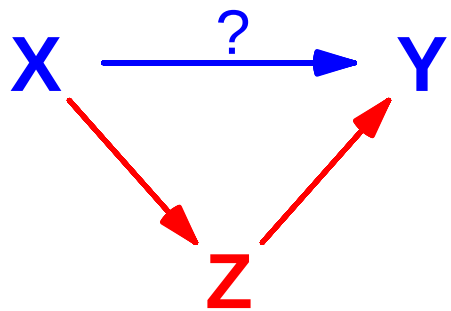
Example: the effect of X and Y on Z:



A simple model may hold: $Z = X + Y + U$,
where U is independent of X and Y.
Hence $Y = Z - X - U$.
We see the association between X and Y only when the "effect" of Z has been taken into account. But this is not the causal effect of X on Y.
One should NOT adjust the analysis for Z!

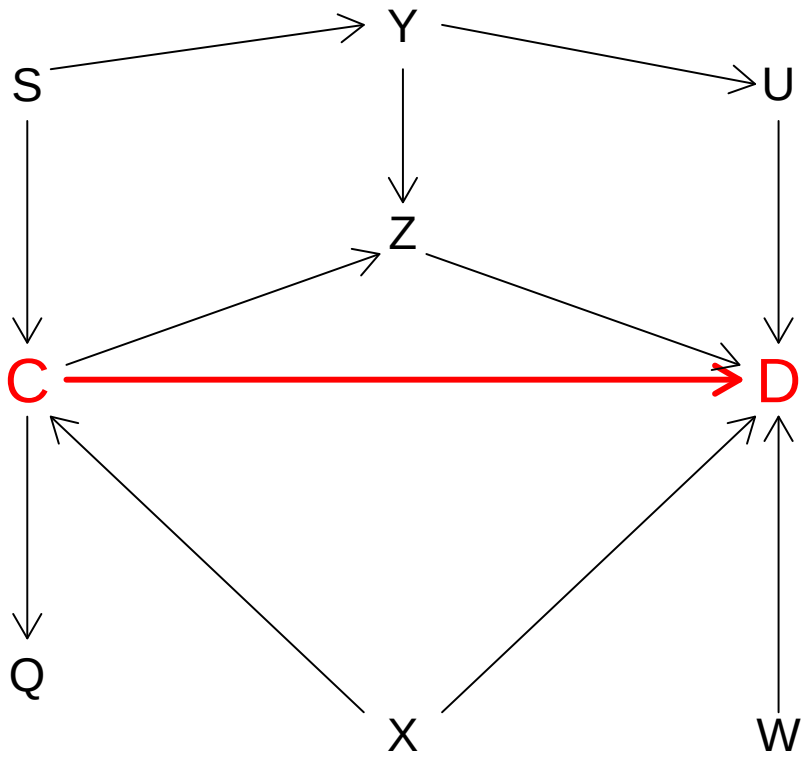
More possibilities: mediation

Example: the effect of X on Y is (partly) mediated by Z:



$Y = X + Z + U$,
If you are interested in the total effect of X on Y – don't adjust for Z!
If you are interested in the direct effect of X on Y – adjust for Z.
(Only if the Z-Y association is unconfounded)

Actually there might be a complicated system of causal effects:



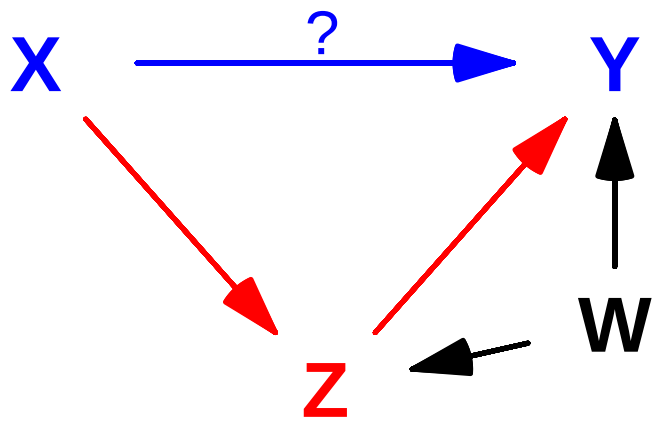
C-smoking; D-cancer

Q, S, U, W, X, Y, Z - other factors that influence cancer risks and/or smoking (genes, social background, nutrition, environment, personality, . . .)

To check for confounding,

1. Sketch a causal graph
2. Remove all arrows corresponding to the causal effect of interest (thus, create a graph where the causal null-hypothesis would hold).
3. Remove all nodes (and corresponding edges) except those contained in the exposure (C) and outcome (D) variables and their (direct or indirect) ancestors.
4. Connect by an undirected edge every pair of nodes that both share a common child and are not already connected by a directed edge.
   - If now C and D are still associated, we say that the $C - D$ association is confounded
   - Identify the set of nodes that need to be deleted to separate C and D – inferences conditional on these variables give unconfounded estimates of the causal effects.

Example: mediation with confounding



Follow the algorithm to show that one should adjust the analysis for W. If W is an unobserved confounder, no valid causal inference is possible in general. However, the total effect of X on Y is estimable.

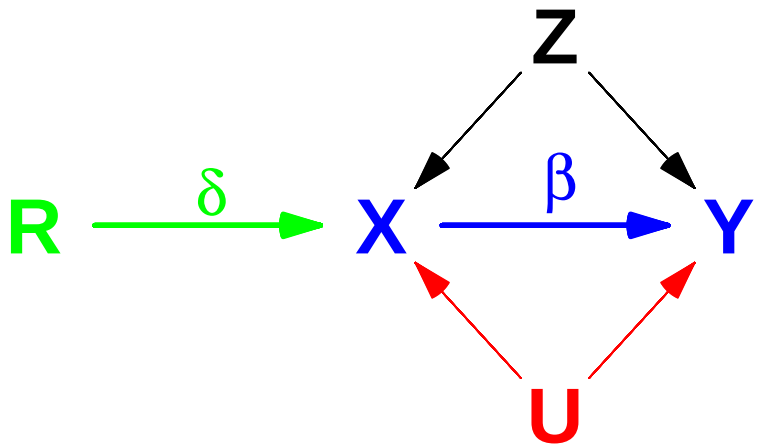
“Mendelian randomization” – genes as Instrumental Variables

- Most of the exposures of interest in chronic disease epidemiology cannot be randomized.
- Sometimes, however, nature will randomize for us: there is a SNP (Single nucleotide polymorphism, a DNA marker) that affects the exposure of interest, but not directly the outcome.
- Example: a SNP that is associated with the enzyme involved in alcohol metabolism, genetic lactose intolerance, etc.

However, the crucial assumption that the SNP cannot affect outcome in any other way than throughout the exposure, cannot be tested statistically!

General instrumental variables estimation

A causal graph with exposure X, outcome Y, confounder U and an *instrument* Z:

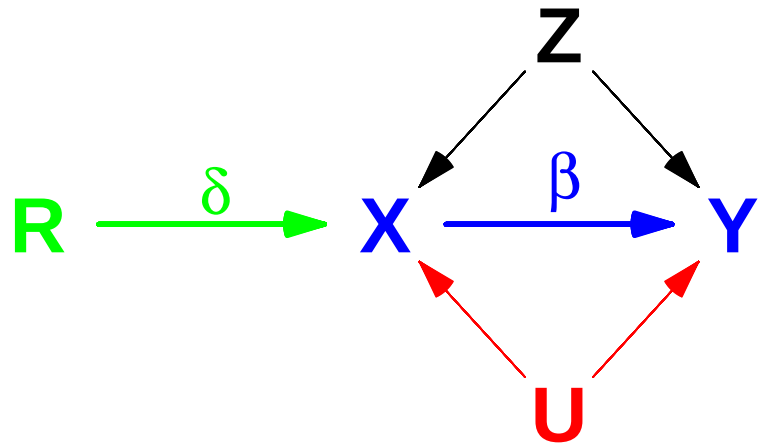


Simple regression will yield a biased estimate of the causal effect of X on Y, as the graph implies:

$$Y = \alpha_y + \beta X + \gamma U + \epsilon,\ \mathrm{E}(\epsilon|X, U) = 0$$

so $\mathrm{E}(Y|X) = \alpha_y + \beta X + \gamma \mathrm{E}(U|X)$.
Thus the coefficient of X will also depend on γ and the association between X and U.

General instrumental variables estimation



$$Y = \alpha_y + \beta X + \gamma U + \epsilon,\ \mathrm{E}(\epsilon|X,U) = 0$$

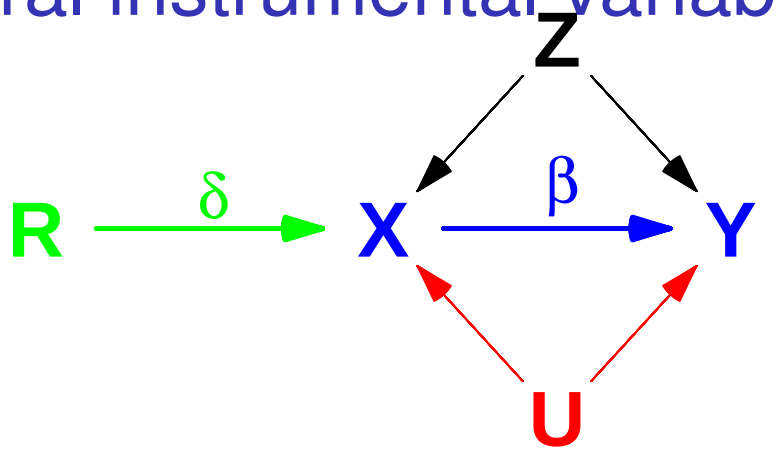
How can Z help?

If $\mathrm{E}(X|Z) = \alpha_x + \delta Z,$ we get

$$\mathrm{E}(Y|Z) = \alpha_y + \beta E(X|Z) + \gamma E(U|Z) = \alpha_y + \beta(\alpha_x + \delta Z) = \alpha_y^* + \beta\delta Z.$$

As δ and $\beta\delta$ are estimable, also β becomes estimable.

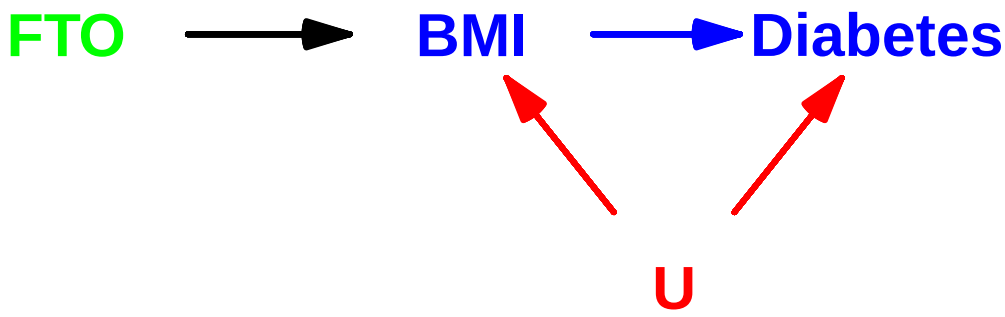
General instrumental variables estimation



1. Regress X on Z, obtain an estimate $\hat{\delta}$
2. Regress Y on Z, obtain an estimate $\hat{\delta\beta}$
3. Obtain $\hat{\beta} = \frac{\hat{\delta\beta}}{\hat{\delta}}$
4. Valid, if Z is not associated with U and does not have any effect on Y (other than mediated by X)
5. Standard error estimation is more tricky – use for instance `library(sem)`, function `tsls()`.

Mendelian randomization example

FTO genotype, BMI and Blood Glucose level (related to Type 2 Diabetes risk; Estonian Biobank, n=3635, aged 45+)



- Average difference in Blood Glucose level (Glc, mmol/L) per BMI unit is estimated as 0.085 (SE=0.005)
- Average BMI difference per FTO risk allele is estimated as 0.50 (SE=0.09)
- Average difference in Glc level per FTO risk allele is estimated as 0.13 (SE=0.04)
- Instrumental variable estimate of the mean Glc difference per BMI unit is 0.209 (se=0.078)

IV estimation in R (using `library(sem)`):

```
> summary(tsls(Glc~bmi, ~fto,data=fen),digits=2)

 2SLS Estimates

Model Formula: Glc ~ bmi

Instruments: ~fto

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-6.3700 -1.0100 -0.0943  0.0000  0.8170 13.2000

             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.210      2.106    -0.6    0.566
bmi            0.209      0.078     2.7    0.008 **
```

IV estimation: can untestable assumptions be tested?

```
> summary(lm(Glc~bmi+fto,data=fen))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.985  0.106   18.75   <2e-16 ***
bmi         0.088  0.004   23.36   <2e-16 ***
fto         0.049  0.030    1.66    0.097 .

For Type 2 Diabetes:
> summary(glm(t2d~bmi+fto,data=fen,family=binomial))
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.515   0.187  -40.18   <2e-16 ***
bmi          0.185   0.006   31.66   <2e-16 ***
fto          0.095   0.047    2.01    0.044 *
```

Does FTO have a direct effect on Glc or T2D?

A significant FTO effect would not be a proof here (nor does non-significance prove the opposite)! (WHY?)

Can we test pleiotropy?

A naïve approach would be to fit a linear regression model for Y, with both X and G as covariates.
But in this case we estimate:

$$\mathrm{E}(Y|X,G) = \text{const} + \beta_{pl}G + \beta X + \gamma \mathrm{E}(U|X,G).$$

It is possible to show that U is not independent of neither X nor G – therefore, the coefficient of G in the resulting model would be nonzero even if $\beta_{pl} = 0$.
Therefore there is no formal test for pleiotropy possible in the case of one genetic instrument – only biological arguments could help to decide, whether assumptions are likelt to be fulfilled
In the case of *multiple genetic instruments* and *meta-analysis*, sometimes the approach of *Egger regression* can be used (Bowden et al, 2015). But even that is not an assumption-free method!

Summary

- There is no unique definition of “the causal effect”
- The validity of any causal effect estimates depends on the validity of the underlying assumptions.
- Adjustment for other available variables may remove (some) confounding, but it may also create more confounding. Do not adjust for variables that may themselves be affected by the outcome.
- Instrumental variables approaches can be helpful, but beware of assumptions!

Some references

- A webpage by Miguel Hernan and Jamie Robins: http://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/
- An excellent overview of Mendelian randomization: Sheehan, N., Didelez, V., Burton, P., Tobin, M., Mendelian Randomization and Causal Inference in Observational Epidemiology, PLoS Med. 2008 August; 5(8).
- A way to correct for pleiotropy bias: Bowden J, Davey Smith G, Burgess S, Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015 Apr;44(2):512-25.
- . . . and how to interpret the findings (warning against overuse): Burgess, S., Thompson, S.G., Interpreting findings from Mendelian randomization using the MR-Egger method, Eur J Epidemiol (2017).