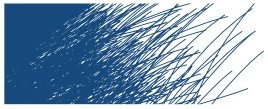


Causal questions and principled answers: a guide through the landscape for practising statisticians

Els Goetghebeur
Ghent University, Belgium
Els.Goetghebeur@UGent.be

for Bianca De Stavola, Saskia Le Cessie, Erica Moodie and Ingeborg Waernbaum

ISCB Vienna, August 26, 2014

Outline

1. Our complete plan - the bigger map
2. Different causal inference approaches
   - Structural (mean) methods with outcome regression and IPW, Instrumental variables, Matching, Mediation analysis, Principal Strata,...
3. Different causal effects targeted:
   - what is potentially changed (direct, indirect,... effects)
   - in what (sub) population
4. Different assumptions made on the data structure
5. Where and when do they overlap and fundamentally differ
6. What is (most) useful/relevant when
7. Discussion

The TG7 broader plan

- I: Target causal effect parameters of different approaches:
  - their interpretation and practical use/relevance
  - the assumptions involved
  - their overlap and distinction
- II: on estimation under the standard assumptions
  - how it is done (incl. software hints)
  - practical properties of the estimators
  - tricks and treats
- III: What it still means when untestable assumptions fail +
  - Clues on failed assumptions
  - Robustness, sensitivity, and the bias-variance trade off
- IV: Missing data
- V: Some guidelines

Links with other topic groups!
descriptives, prediction, missing data...

The TG7 plan - our approach

- work from simple to complex
- from binary trt. to continuous and static or dynamic treatment regimes over time
- from binary over continuous, right censored survival to generally repeated outcomes over time
- from (semi)-parametric to more flexible prediction models
- from (repeated) 'cross-sectional' to longitudinal data set-up, prospective to retrospective designs, ...
- population constant effects and exposures interacting; conditional and average effects
- acknowledging increasing levels of (unmeasured) confounding
- handling missing data

Pointers to tutorials and software implementation

Worked out case studies , simulation studies

From paper(s) to website with links: getting more people involved

Different approaches & own targeted causal effects on Y

- Exposures:
  Assigned treatment A $\leftarrow$ policy, scientists, caregiver
  Manifested treatment M $\leftarrow$ patient, patient management
  (Think Statin versus Non-statin use)
- Interventions:
  Single $(\mathfrak{a})$ intervention $->$ total effect
  Double intervention $(\mathfrak{a}, \mathfrak{m})$ $->$ direct and indirect effects
- Effect measures:
  Marginal versus Baseline (L) stratified mean effect
- Target population: Post treatment stratified?
  ITT, AT, PP, PS, TAT, extrapolation, other ...

Causal effect of possible exposure on potential outcome

Causal: action/decision set in 2(+) directions $\mathfrak{a} = 1$ or $\mathfrak{a} = 0$.
Learn about the expected consequences of our choice/decision.
Causal inference: evidence to support this decision, from data.

Possible action in calligraphics like $\mathfrak{a} = 1$ for set action level to 1.
Consequences: potential outcomes: $Y(\mathfrak{a} = 1)$ or $Y(1)$, defined $\forall$.
One potential outcome $Y(\mathfrak{a})$ seen (Y) in subset
$\{Y|A = 1\} = \{Y(1)|A = 1\}$

- $E(Y|A = a) = E(Y(a)|A = a)$

Causal effect of possible exposure on potential outcome

Causal: action/decision set in 2(+) directions $\mathfrak{a} = 1$ or $\mathfrak{a} = 0$.
Learn about the expected consequences of our choice/decision.
Causal inference: evidence to support this decision, from data.

Possible action in calligraphics like $\mathfrak{a} = 1$ for set action level to 1.
Consequences: potential outcomes: $Y(\mathfrak{a} = 1)$ or $Y(1)$, defined $\forall$.
One potential outcome $Y(\mathfrak{a})$ seen (Y) in subset
$\{Y|A = 1\} = \{Y(1)|A = 1\}$

- $E(Y|A = a) = E(Y(a)|A = a)$ and
- for baseline characteristics ℓ:
  $E(Y|A = a, L = \ell) = E(Y(a)|A = a, L = \ell)$

Causal effect of possible exposure on potential outcome

Causal: action/decision set in 2(+) directions $\mathfrak{a} = 1$ or $\mathfrak{a} = 0$.
Learn about the expected consequences of our choice/decision.
Causal inference: evidence to support this decision, from data.

Possible action in calligraphics like $\mathfrak{a} = 1$ for set action level to 1.
Consequences: potential outcomes: $Y(\mathfrak{a} = 1)$ or $Y(1)$, defined $\forall$.
One potential outcome $Y(\mathfrak{a})$ seen (Y) in subset
$\{Y|A = 1\} = \{Y(1)|A = 1\}$

- $E(Y|A = a) = E(Y(a)|A = a)$ and
- for baseline characteristics ℓ:
  $E(Y|A = a, L = \ell) = E(Y(a)|A = a, L = \ell)$

  $= E(Y(a)|L = \ell)$ *if* $Y(\mathfrak{a}) \coprod A|L \quad \forall \mathfrak{a}.$

  'No unmeasured confounding'

Two 'Universal' Assumptions:

UA1 **No interference between subjects** a subject's potential outcome is not influenced by the treatment received by others

UA2 **The intervention is well-defined** so that observed and potential outcomes coincide when their action levels are identical

And two *Adjustment Assumptions:*

AA1 **No unmeasured confounding** The conditional probability of receiving the treatment depends only on measured covariates, and not on any unmeasured covariate.

AA2 **Positivity** The conditional probability of receiving the treatment is neither zero nor one

Outcome regression

$$Y(\mathfrak{a}) \coprod A|L \quad \forall \mathfrak{a} \quad \Rightarrow$$

$$\{Y|L, A = a\} = \{Y(a)|L, A = a\} \stackrel{d}{=} \{Y(a)|L\}$$

Hence simply regress Y on L in several A-defined strata
to infer the population distribution of $Y(a)$ conditional on L.

regress Y on L in Statin users	$-> f_1(y\|\ell)$
regress Y on L in Non-statin users	$-> f_0(y\|\ell)$

Outcome regression

$Y(a) \coprod A|L \quad \forall a \quad \Rightarrow$

$$\{Y|L, A = a\} = \{Y(a)|L, A = a\} \stackrel{d}{=} \{Y(a)|L\}$$

Hence simply regress Y on L in several A-defined strata
to infer the population distribution of $Y(a)$ conditional on L.

regress Y on L in Statin users	$-> f_1(y\|\ell)$
regress Y on L in Non-statin users	$-> f_0(y\|\ell)$

Challenges:

- With 'high' dimension of ℓ : confidence in a correct model
- $L-$distribution for (non)treated does not overlap $(\pm)$
  e.g. in the young and fit you may find no statin users
- $E(Y|L, A = 1) - E(Y|L, A = 0) =$
  $E(Y(1)|L) - E(Y(0)|L) = \psi(L)$ i.e. may differ over L

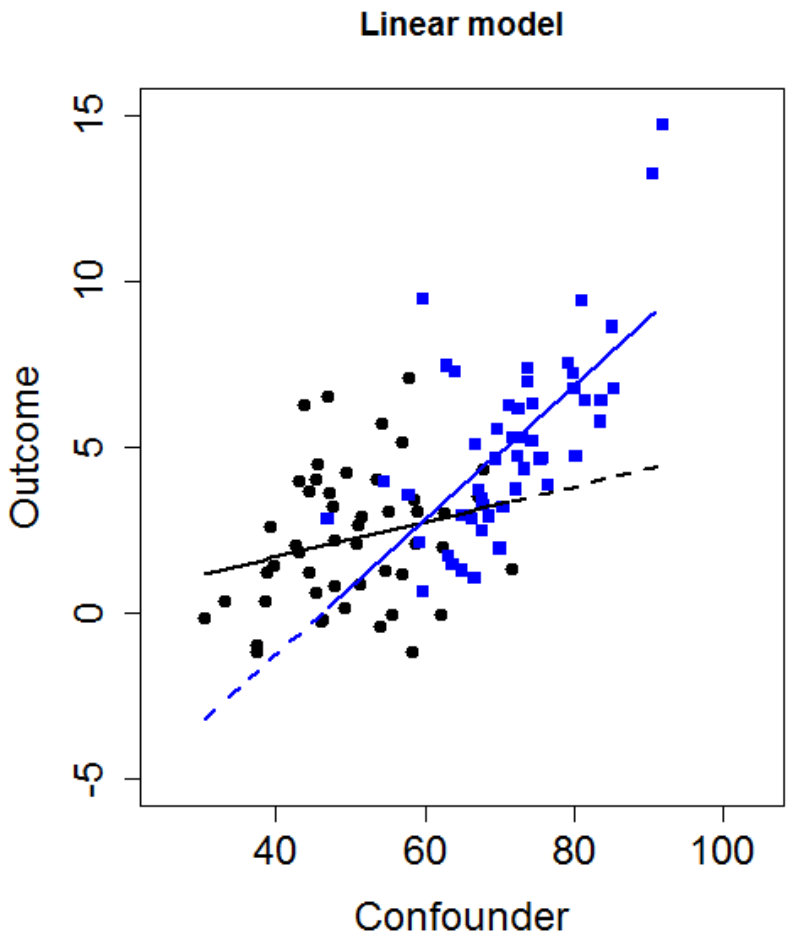
Challenges:

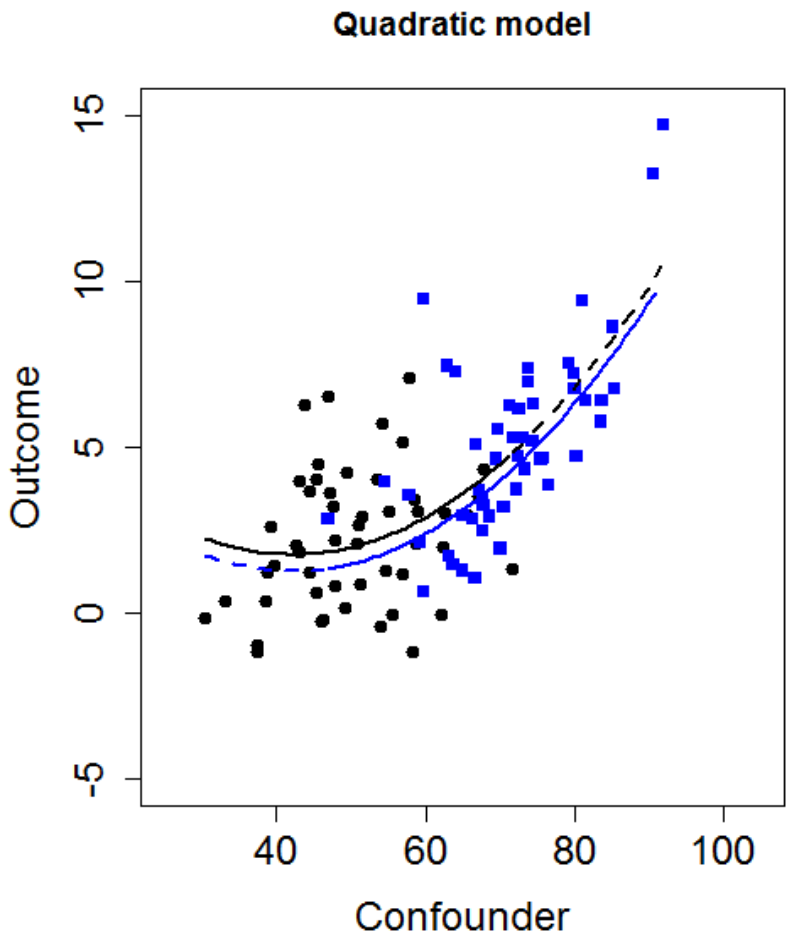
- With 'high' dimension of ℓ : confidence in a correct model
  $->$ a number of propensity score $\pi_a(L)$ based solutions:
  regress, stratify, match, inverse weighting, DR
  all involve a regression model for observed action $\pi_a(L)$,
  regressing A on L.
- $L-$distribution for non- treated does $(\pm)$ not overlap
  e.g. in the young and fit you may find no statin users $\Rightarrow$
  $->$ restrict target population to the common L-space
  (otherwise positivity violated)
  outcome regression may hide this and extrapolate, also IPW
  unless 'close to zero'

Challenges:

- With 'high' dimension of ℓ : confidence in a correct model
  $->$ a number of propensity score $\pi_a(L)$ based solutions:
  regress, stratify, match, inverse weighting, DR
  all involve a regression model for observed action $\pi_a(L)$,
  regressing A on L.
- $L-$distribution for non- treated does $(\pm)$ not overlap
  e.g. in the young and fit you may find no statin users $\Rightarrow$
  $->$ restrict target population to the common L-space
  (otherwise positivity violated)
  outcome regression may hide this and extrapolate, also IPW
  unless 'close to zero'
- $E(Y|L,A=1)-E(Y|L,A=0)=E(Y(1)|L)-E(Y(0)|L)$
  may differ over L : $\psi(\ell)$.
  $->$ model this function of L, then average over L to obtain
  ATE: the population average treatment effect
  $\neq$ ATAT: average treatment effect among the treated

Outcome regression hiding extrapolation





What set of confounders L?

$Y(\mathfrak{a}) \coprod A|L \quad \forall \mathfrak{a}.$

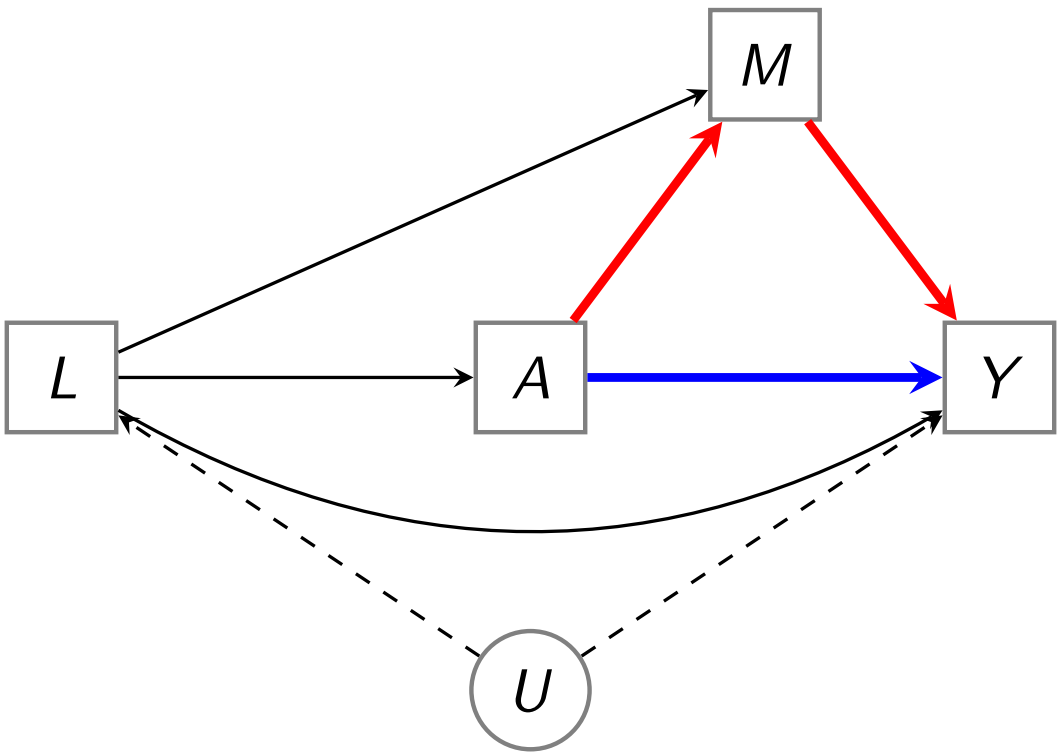
'No unmeasured confounding'

- Not unique, the set L satisfying 'no unmeasured confounding'
- Augmenting and reducing the set L can lead to violated assumption
- Adding a variable can turn a non-confounder into confounder
- Adding a variable can make an extisting confounder redundant

Chosen covariates + functional form =
correct Outcome Regression and Propensity Score models.

Mediation: Assigned versus Manifested treatment

Experiment: trt assigned, A, may differ from manifested, M :
Instrumental variables, principal strata, mediation analysis



Direct effect and indirect effects with many definitions...

Direct and indirect effects

- *Controlled Direct Effect* of A on Y with M controlled at m

$$CDE(m) = E\{Y(1,m)\} - E\{Y(0,m)\}$$

Direct and indirect effects

- *Controlled Direct Effect* of A on Y with M controlled at m

$$CDE(m) = E\{Y(1,m)\} - E\{Y(0,m)\}$$

$m = 0$: Promised treatment withheld, or
$m = 1$: Trt also obtained without prescription/reimbursement

Direct and indirect effects

- *Controlled Direct Effect* of A on Y with M controlled at m

$$CDE(m) = E\{Y(1,m)\} - E\{Y(0,m)\}$$

$m = 0$: Promised treatment withheld, or
$m = 1$: Trt also obtained without prescription/reimbursement

- *Pure Natural Direct Effect* of A on Y

$$PNDE = E\{Y(1,M(0))\} - E\{Y(0,M(0))\}.$$

Direct and indirect effects

- *Controlled Direct Effect* of A on Y with M controlled at m

$$CDE(m) = E\{Y(1,m)\} - E\{Y(0,m)\}$$

$m = 0$: Promised treatment withheld, or
$m = 1$: Trt also obtained without prescription/reimbursement

- *Pure Natural Direct Effect* of A on Y

$$PNDE = E\{Y(1,M(0))\} - E\{Y(0,M(0))\}.$$

Estimating a placebo effect when $M(0) = 0$,
reimbursement/supporting effect when $M(0) = 1$.

Direct and indirect effects

- *Controlled Direct Effect* of A on Y with M controlled at m

$$CDE(m) = E\{Y(1,m)\} - E\{Y(0,m)\}$$

- *Pure Natural Direct Effect* of A on Y

$$PNDE = E\{Y(1,M(0))\} - E\{Y(0,M(0))\}.$$

- *Total Natural Indirect Effect*

$$TNIE = TCE - PNDE = E\{Y(1,M(1))\} - E\{Y(1,M(0))\}.$$

Average effect of 'Assigned treatment and 'get it versus not get it, among compliers ', 'no manifest change among others '.

Principal Strata

Possible assignments $\mathfrak{a} = 1/0$ translate into potentially observed: $(\mathfrak{a} = 0, M(0), Y(0, M(0))$ and $(\mathfrak{a} = 1, M(1), Y(1, M(1)))$.

Principal strata conceive joint manifestations of trt. $(M(0), M(1))$

N	Never treated	$[M(0) = 0, M(1) = 0]$
C	Compliers	$[M(0) = 0, M(1) = 1]$
D	Defiers	$[M(0) = 1, M(1) = 0]$
A	Always treated	$[M(0) = 1, M(1) = 1]$

Estimating ITT per stratum:

PRO : Average effect of assignment explained by manifest trts
Challenged use : strata not identified/identifiable
Assignment in trial may differ from future prescription impact.

Suppose we did have/know it all

S	baseline risk	prev.	trt	$\mathfrak{m} = 0$	$E(Y(1, M(1)) - Y(0, M(0))\|S)$	effect assd.
N	high	π_N	Never	α_N	ψ_N	$\equiv 0$?
C	medium	π_C	Compliers	α_C	ψ_C	
D	very high	π_D	Defiers	α_D	ψ_D	$\equiv -\psi_C$?
A	very low	π_A	Always	α_A	ψ_A	$\equiv 0$?

Last column assumes effect of manifest treatment only.

ITT, As Treated and Per Protocol IF A is randomized

- Total Causal effect of assigned treatment A equals ITT and can be estimated as: $E\left(E(Y|A=1,L) - E(Y|A=0,L)\right)$ or $\sum_{s=1}^{4} \pi_s E(Y(1,M(1)) - Y(0,M(0))|S=s)$.

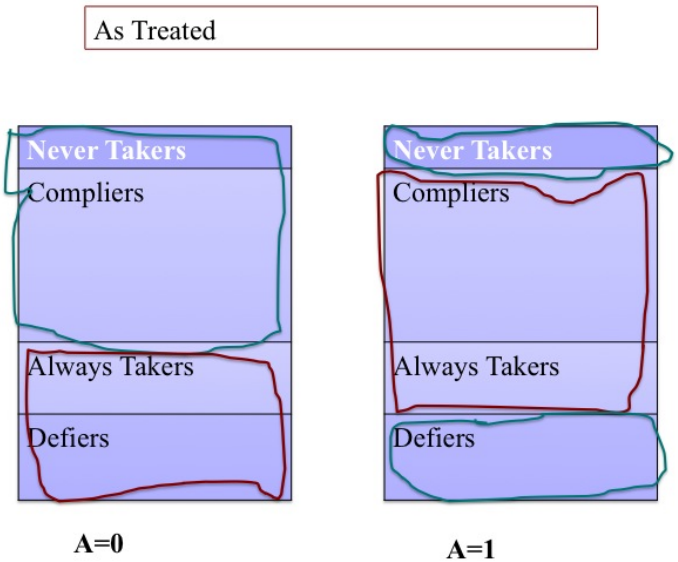
ITT, As Treated and Per Protocol IF A is randomized

- Total Causal effect of assigned treatment A equals ITT and can be estimated as: $E\left(E(Y|A=1,L)-E(Y|A=0,L)\right)$ or $\sum_{s=1}^{4}\pi_s E(Y(1,M(1))-Y(0,M(0))|S=s)$.

$$\begin{aligned} \text{As Treated Effect} &= E(Y|M=1)-E(Y|M=0) \\ &= \text{ITT for } \{\text{compliers}\} \\ &\text{-} \quad \text{ITT for } \{\text{defiers}\} + \end{aligned}$$

  - never takers risk contributes twice to control group
  - always takers risk contributes twice to treatment group

  the full population is involved .



$$E(Y|M=1) - E(Y|M=0) =$$
$$\pi_C\psi_C - \pi_D\psi_D + 2\pi_A\alpha_A - 2\pi_N\alpha_N$$

- If < 0 this would reveal that those who are on the treatment as a group are lower risk than those who are off it as a group.
- In the absence of Never and Always Takers (or if this term otherwise cancels out- which is unlikely) $->$may reveal that you are better off taking it than not if you can (i.e. causal)

Per Protocol Effect = ITT for { compliers } +

- Never Takers risk contributes to control group
- Always Takers risk contributes to treatment group
- Defiers are not considered

In the randomized trial

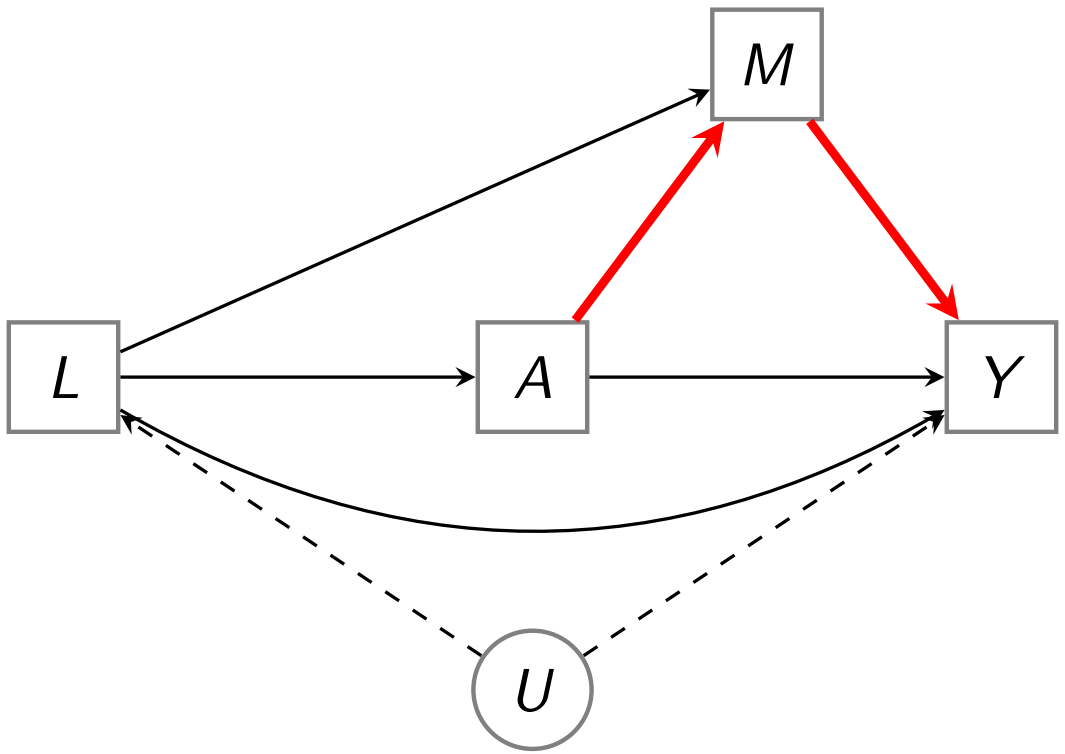
If experimental treatment not available outside the trial:
we have no Defiers and no Always Takers.

With only compliers and never takers:

- percentage of compliers easily found in the control arm.
- if drug only available on prescription: only these subjects will stay on the drug and the ITT for them is the most relevant
- hence % compliers + complier ITT effect is relevant measure for policy makers/prescribers and patients
- Follow-up work on the never takers is needed: what can they get that helps them?

Baseline confounding

Indirect effect

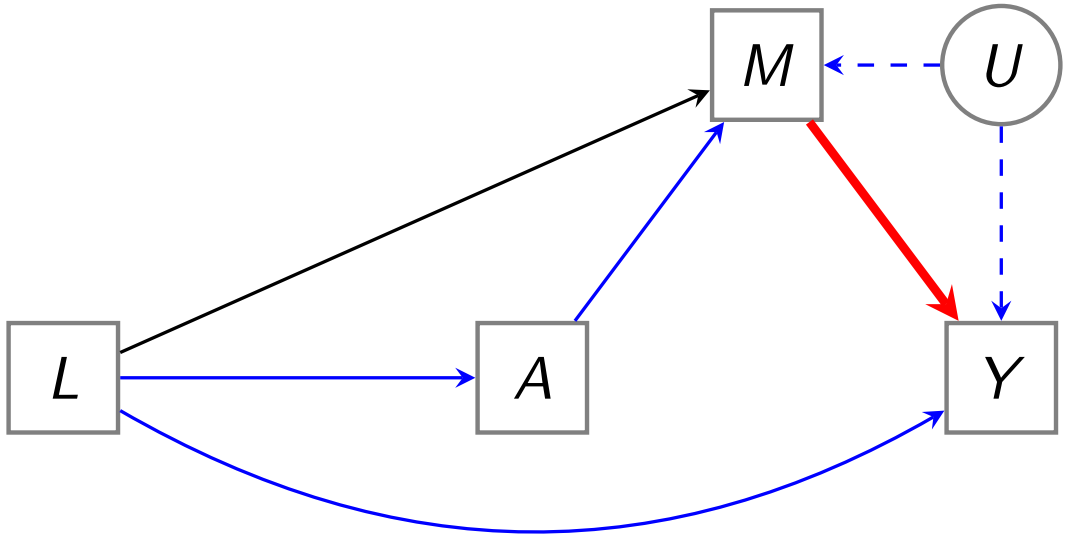


Following assumptions are made:

(A1) No unmeasured confounders of A-Y relationship

(A2) No unmeasured confounders of M-Y relationship

A as an instrumental variable



Targeted: causal effect of M on Y

Assume:

- No direct effect of A on Y
- but (strong) effect of A on M desired

Unmeasured confounders U of M and Y allowed for.

Causal effect of ($\mathfrak{m} = 1$ versus $= 0$) among the treated $M = 1$.

Danaei et al., 2011, SMMR

Observational study with repeated measures of 'on statins' or not.
An 'ITT like' analysis targets effect of statin initiation
in population of non-statins-takers for the past 2 years.
outcome Y = time to occurrence of CHD/death or LOF/censoring.

At t_0 (Jan. 2000) $M = 1(0)$ for those who *initiate* statins (or not)
Conditional on L, statin initiation is assumed random

- Y of initiating group versus the non-initiating group, given L ignores statin use in the months to come $\Rightarrow$ 'ITT' .
- 'Per protocol analysis' : considers continued statin use versus continued non-use, subjects censored when off protocol.
- 'As treated:' instantaneous risk depends on time-varying history of 'total duration of treatment so far'

Strong confounding by indication: high risk patients more likely to take statins $\Rightarrow$ residual confounding?

Discussion

- All V phases: a looong term project, we will need help (comment/contribute)
- Literature is fast growing, working in 'strata'. We wish to provide basic entrance map with directions and anchor points.
- Phase I: 'What question are we answering by distinct principled approaches' and what do we want ?
  - Crucial starting point in our view
  - Not trivial
  - Often ill understood and overlooked by users of available technologies: at the abstract as well as specific level
- "An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem" ? John Tukey

- Causal inference in epidemiology is better viewed as an exercise in measurement of an effect rather than as a criterion-guided process for deciding whether an effect is present or not. (Rothman and Greenland, 2005)