Targeted Learning for Data Adaptive Causal Inference in Observational and Randomized Studies

Maya Petersen and Mark van der Laan

Department of Biostatistics, University of California, Berkeley School of Public Health

Part III: From causal questions to the statistical estimation problem

Extension to multiple/longitudinal interventions

Outline

- Longitudinal casual models and causal graphs
- Counterfactual casual parameters to summarize the joint effects of multiple interventions
- Identification in the longitudinal setting: The challenge of time-dependent confounding

Example: Abacavir and Cardiovascular Disease

- Analysis of observational data from several cohorts suggested abacavir use associated with increased risk of myocardial infarction among treated HIV-infected population
  - Other analyses found no evidence of such an association….
- Example of a causal question: Does current use of abacavir (ABC) increase risk of myocardial infarction (MI)?

Notation for Longitudinal Data

- L(t)= covariates at time t, t=1,…,K+1
  - The time-varying equivalent of W
  - As usual, a node can be multi dimensional
- Y(t)= outcome at time t, t=1,…,K+1
  - Sometimes defined as a subset of L(t)
  - Alternative: Y measured only at the end of follow up, sometimes defined as a subset of L(K+1)
- A(t)=exposure/treatment at time t, t=1,…,K

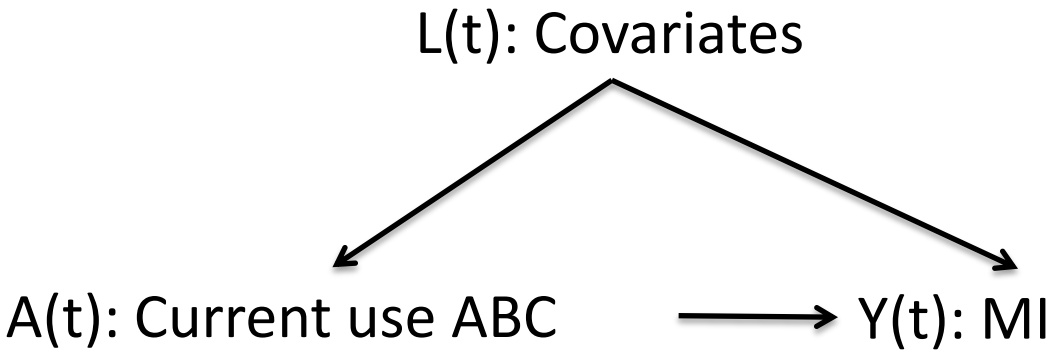
Example: Effect of current abacavir use on MI risk

- Monthly Data (Time in month increments)
- A(t)=Indictor current abacavir use at start of month t
- Y(t)=Indicator MI by close of the month t
- L(t)=Covariates in prior month
  - Other Drugs, Lipids, DM, HTN…
  - This can include summaries of patient history up to start of the month, including past CHD
- O(t)=(L(t),A(t),Y(t)), t=1,…,K

Example: Effect of current abacavir use on MI risk?

- Structural Causal Model/Graph for a single time point?

$$L(t) = f_{L(t)}(U_{L(t)})$$
$$A(t) = f_{A(t)}(L(t), U_{A(t)})$$
$$Y(t) = f_{Y(t)}(L(t), A(t), U_{Y(t)})$$



Example: Effect of current abacavir use on MI risk?

- Counterfactual outcomes: $Y_{a(t)}(t)$, t=1,…,K
  - $Y_1(t)$: counterfactual MI status at end of month if used abacavir in month t
  - $Y_0(t)$: counterfactual MI status at end of month if did not use abacavir in month t
- Possible target causal quantity
  - $E(Y_1(t)-Y_0(t)|Y(t-1)=0)$
  - Difference in risk of incident MI in month t if did vs. did not use abacavir

Conditioning on survival under no intervention

Example: Effect of current abacavir use on MI risk?

- For a given time point, the data are analogous to the (W,A,Y) data we have been discussing
  - We can consider this as a repeated point treatment data structure
  - Allows us to use Model, Data, Identifiability Result, and Estimators previously introduced
  - Cross-Validation and inference need to respect repeated measures data structure
    - Specify patient ID as unit of independence

Cumulative effects of longitudinal treatments?

- What if we want to know about the effects of cumulative exposure to abacavir?
  - Alternative target parameter that investigates the effect of extended abacavir use patterns?
- Need to go beyond repeated point treatment formulation
  - SCM that incorporates time-varying covariates and time-varying treatment
  - Counterfactual outcomes indexed by interventions on more than one treatment node

SCM for Longitudinal Data

- Over-bars used to refer to the history of a variable $\bar{A}(t) = \{A(1), A(2), ..., A(t)\}$

  $\bar{L}(t) = \{L(1), L(2), ..., L(t)\}$

- A common SCM: Assumes each variable may be affected by all preceding variables

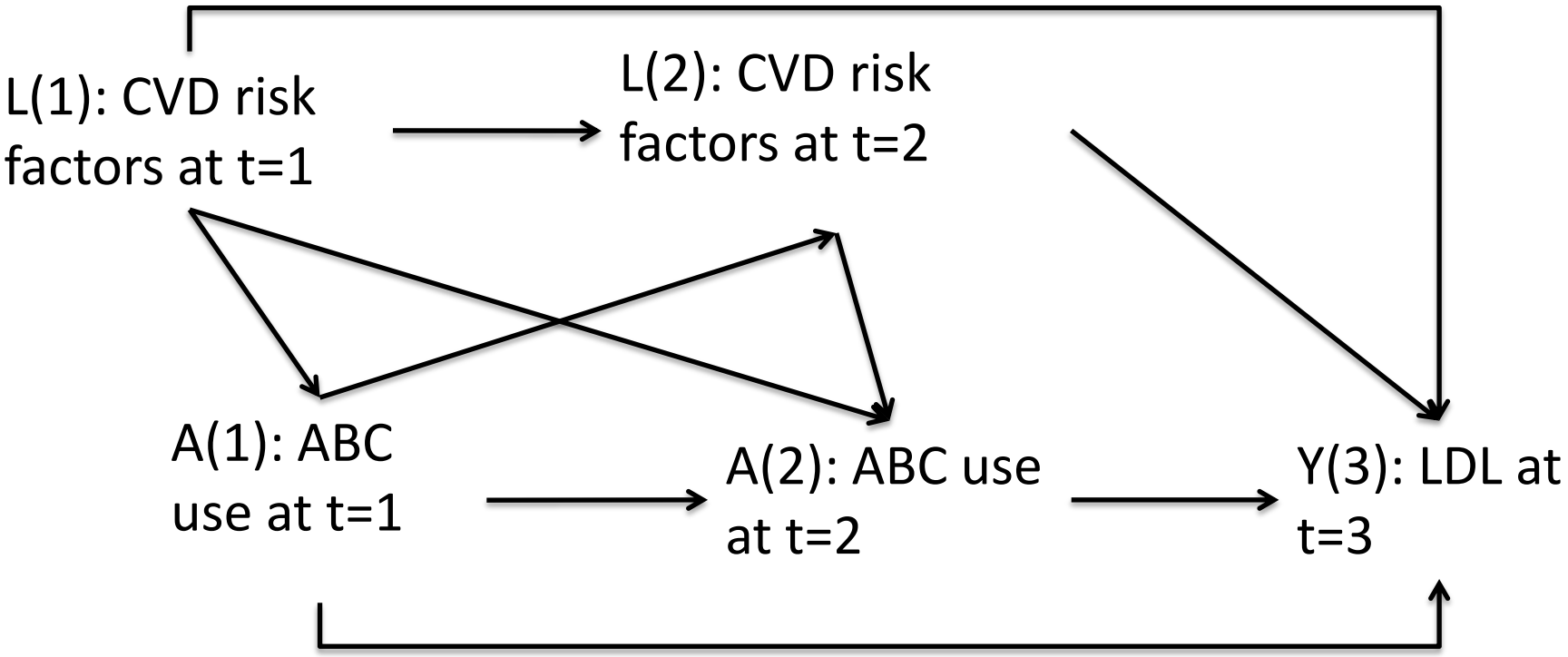
$$L(1) = f_{L(1)}(U_{L_1})$$
$$A(1) = f_{A(1)}(L(1), U_{A(1)})$$
$$L(t) = f_{L(t)}(\bar{A}(t-1), \bar{L}(t-1), U_{L(t)}), t = 2, ..., K+1$$
$$A(t) = f_{A(t)}(\bar{A}(t-1), \bar{L}(t), U_{A(t)}), t = 2, ..., K$$

Simplified Abacavir Example

- Say we measure
  - CHD risk factors (including lipids) at t=1 and t=2
  - Abacavir use at t=1 and t=2
  - Outcome= LDL cholesterol at t=3
  - Assume no deaths, censoring, or missing data for now
- We are interested in the difference in expected LDL at t=3 if
  - all subjects had used abacavir at t=1 and t=2

    versus
  - no subjects had used abacavir at t=1 and t=2

Abacavir Example: Longitudinal Causal Graph



Counterfactuals indexed by longitudinal exposures

- Original SCM

$$L(1) = f_{L(1)}(U_{L_1})$$
$$A(1) = f_{A(1)}(L(1), U_{A(1)})$$
$$L(2) = f_{L(2)}(L(1), A(1), U_{L(2)})$$
$$A(2) = f_{A(2)}(A(1), \bar{L}(2), U_{A(2)})$$
$$Y = f_Y(\bar{L}(2), \bar{A}(2), U_Y)$$

- Modified SCM, intervening on abacavir use at times 1 and 2?

Counterfactuals indexed by longitudinal exposures

- Original SCM

$$L(1) = f_{L(1)}(U_{L_1})$$
$$A(1) = f_{A(1)}(L(1), U_{A(1)})$$
$$L(2) = f_{L(2)}(L(1), A(1), U_{L(2)})$$
$$A(2) = f_{A(2)}(A(1), \bar{L}(2), U_{A(2)})$$
$$Y = f_Y(\bar{L}(2), \bar{A}(2), U_Y)$$

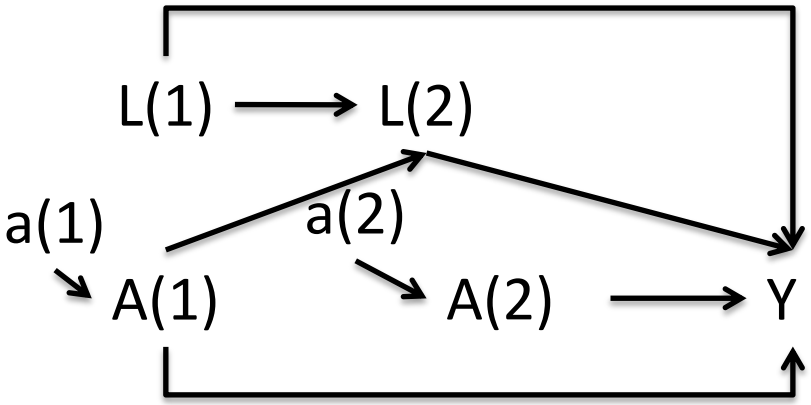
- Modified SCM, intervening on abacavir use at times 1 and 2

$$L(1) = f_{L(1)}(U_{L_1})$$
$$A(1) = a(1)$$
$$L(2) = f_{L(2)}(L(1), a(1), U_{L(2)})$$
$$A(2) = a(2)$$
$$Y = f_Y(\bar{L}(2), \bar{a}(2), U_Y)$$

Counterfactuals indexed by longitudinal exposures

- Modified SCM/Graph

$$L(1) = f_{L(1)}(U_{L_1})$$
$$A(1) = a(1)$$
$$L(2) = f_{L(2)}(L(1), a(1), U_{L(2)})$$
$$A(2) = a(2)$$
$$Y = f_Y(\bar{L}(2), \bar{a}(2), U_Y)$$



- Defines counterfactual outcome intervening on ABC use at two time points:

$$Y_{a(1),a(2)} = Y_{\bar{a}}$$

Intervention on counterfactual exposure history

Abacavir Example: Defining a longitudinal target parameter

- Question: How would expected LDL at t=3 have differed if all subjects had used abacavir at t=1 and t=2 versus if all subjects had not used abacavir at t=1 and t=2 ?
- How would you write the corresponding target causal parameter?

$$E(Y_{11} - Y_{00})$$

Defining target causal quantity using a Longitudinal Marginal Structural Model

- Example: How does cumulative time exposed to abacavir affect LDL at the end of the study?
  - Ex. working MSM $E(Y_{\bar{a}}) = \beta_0 + \beta_1 \sum_{t=1}^{K} a(t)$
- How does this effect differ depending on baseline renal function (V)?
  - Ex. Working MSM

$$E(Y_{\bar{a}}|V) = \beta_0 + \beta_1 \sum_{t=1}^{K} a(t) + \beta_2 V + \beta_3 V \times \sum_{t=1}^{K} a(t)$$

Survival Data

- So far, we have focused on a continuous outcome, measured at the end of the study on everybody (assumed no death or censoring/LTFU)
- Now let's return to the original outcome: MI
  - Y(t)= indicator of MI by end of month t

Examples of target causal quantities with survival outcome

- Example: How does counterfactual (discrete) hazard of MI vary as a function of cumulative abacavir exposure since study enrollment?

$$P(Y_{\bar{a}}(t) = 1 | Y_{\bar{a}}(t-1) = 1)$$

- Example of MSM we could use to define the target quantity?

$$logit\left(P(Y_{\bar{a}}(t) = 1 | Y_{\bar{a}}(t-1) = 0)\right) =$$

$$\beta_0 + \beta_1 t + \beta_2 \sum_{j=1}^{t} a(j) + \beta_3 t \times \sum_{j=1}^{t} a(j)$$

What about censoring?

- So far, we have assumed no censoring/loss to follow up
  - All subjects followed until min(K+1, event time)
- In practice, generally not true
  - Abacavir example- data are gathered as part of (several) clinical cohorts
    - Patients transfer to other clinics, drop out of care…
  - Loss to follow up ubiquitous in both observational and RCT datasets

Incorporating censoring

- We can incorporate censoring in the SCM as a set of an additional X nodes in our graph (with their own structural equation)
- C(t)= Indicator still in follow up at time t

$$L(t) = f_{L(t)}(\bar{L}(t-1), \bar{C}(t-1), \bar{A}(t-1), U_{L(t)}), t = 1, ..., K+1$$
$$C(t) = f_{C(t)}(\bar{L}(t), \bar{C}(t-1), \bar{A}(t-1), U_{C(t)}), t = 1, ..., K$$
$$A(t) = f_{A(t)}(\bar{L}(t), \bar{C}(t), \bar{A}(t-1), U_{A(t)}), t = 1, ..., K$$
$$Y(t) \subset L(t)$$

Defining a target causal quantity in the presence of censoring

- Can now think of intervening not only on exposure/treatment at multiple time points, but also intervening on censoring/loss to follow up
- Example: What is the effect of cumulative abacavir exposure on hazard of MI *if all loss to follow up from the cohort had been prevented*?

Defining a target causal quantity in the presence of censoring

- Counterfactuals of interest defined by intervening on two types of nodes:
  - Exposure (abacavir use up till time t)
  - Censoring (stay in cohort up till time t)

$$Y_{\bar{a},\bar{c}=0}(t), \bar{a} \in \mathcal{A}, t = 1, ..., K+1$$

$$L(t) = f_{L(t)}(\bar{L}(t-1), \bar{C}(t-1) = 0, \bar{a}(t-1), U_{L(t)}), t = 1, ..., K+1$$
$$C(t) = 0, t = 1, ..., K$$
$$A(t) = a(t), t = 1, ..., K$$
$$Y(t) \subset L(t)$$

Example of target causal quantities with survival outcome and censoring

- Discrete counterfactual hazard:

$$P(Y_{\bar{a},\bar{c}=0}(t) = 1 | Y_{\bar{a},\bar{c}=0}(t-1) = 0)$$

- Again, can pose a (working) MSM for how this varies as a function of time and cumulative exposure

$$P(Y_{\bar{a},\bar{c}=0}(t) = 1 | Y_{\bar{a},\bar{c}=0}(t-1) = 0) = m(\bar{a}, t|\beta)$$

Additional target causal quantities: Effects of Dynamic Regimes

- Static regime: Set each intervention node equal to some constant
  - Irrespective of subject characteristics
  - Ex: Always use abacavir
- Dynamic regime: A subject-responsive strategy for assigning treatment
  - Assign a value to each intervention node based on some known function of the observed past

Effects of Dynamic Regimes

- Ex. Dynamic regime
  - Always use abacavir **unless** a contraindication (CI) develops, in which case switch to other drug
  - Ie set Abacavir use according to rule d_t(CI(t)):

$$d_t(CI(t)) = 1 \text{ if } CI(t) = 0$$
$$= 0 \text{ if } CI(t) = 1$$

- Effects of dynamic regimes can be defined analogously to effects of static treatment regimens
  - Ex: $E(Y_{\bar{d}}(t) - Y_{\bar{0}}(t)),$

where $\bar{d} = d_1(CI(1)), d_2(CI(2)), \ldots, d_t(CI(t))$

Dynamic Marginal Structural Models

- Dynamic regime might also be indexed by some threshold θ
  - Ex. Don't use abacavir (ie use alternative such as tenofovir) **unless renal function falls below some value** θ, in which case switch to abacavir
  - I.e. set Abacavir use according to rule d_θ(CI):

$$d_\theta(RF(t)) = 0 \text{ if } RF(t) \geq \theta$$
$$= 1 \text{ if } RF(t) < \theta$$

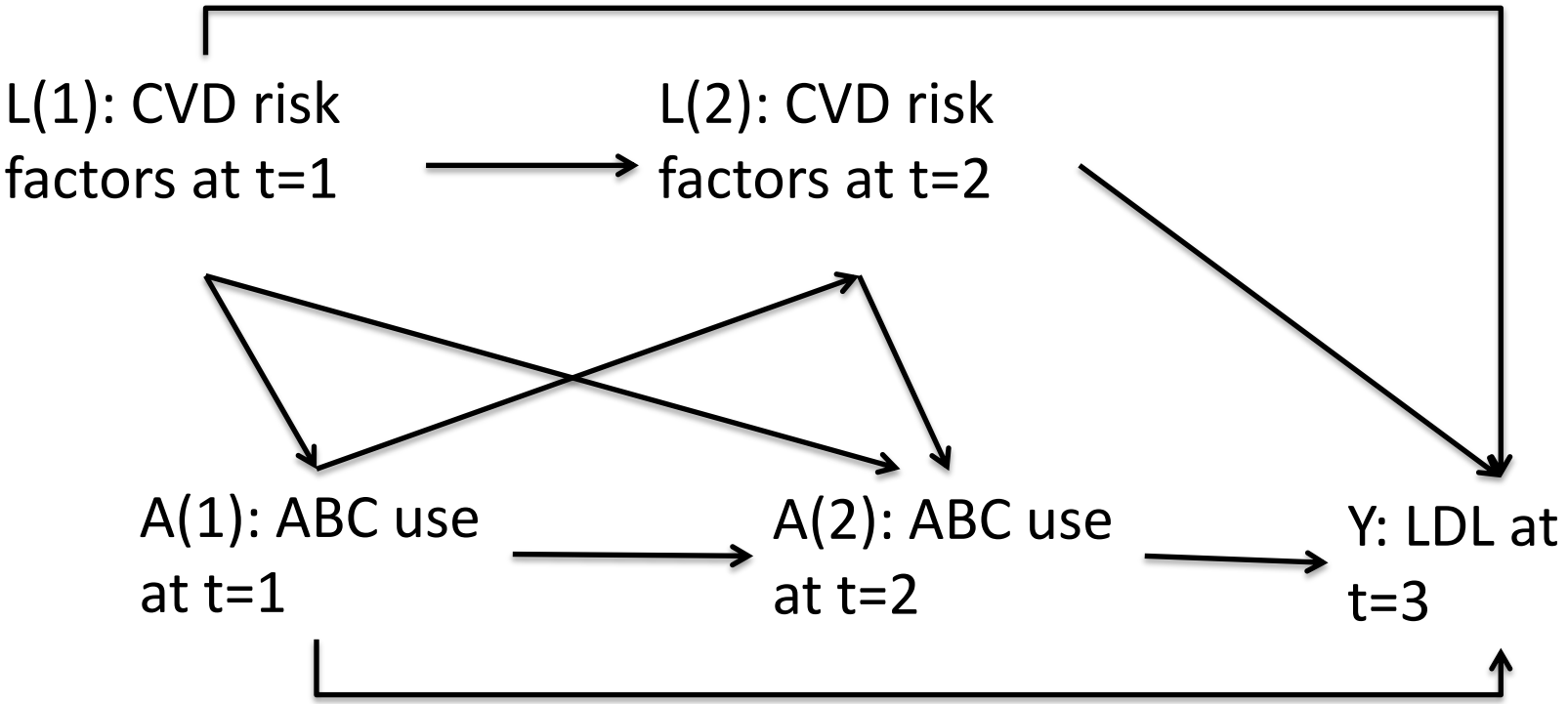
- MSM can be used to summarize how expected counterfactual outcome varies as a function of θ
  - Ex: $E(Y_{\bar{d}_\theta}) = m(\theta|\beta)$

Identifiability for longitudinal exposures

- What causal assumptions are sufficient for our target longitudinal causal parameter to be identified as a parameter of the observed data distribution?

- Back to our simplified example for illustration
  - Effect of Abacavir use at t=1 and t=2 on LDL at t=3
  - Measure CVD risk factors at t=1 and t=2
  - Assume no deaths, censoring, or missing data

Abacavir Example: SCM/Graph



$$L(t) = f_{L(t)}(\bar{L}(t-1), \bar{A}(t-1), U_{L(t)}), t = 1, 2, 3$$

$$A(t) = f_{A(t)}(\bar{L}(t), \bar{A}(t-1), U_{A(t)}), t = 1, 2$$

$$Y \subset L(3)$$

Abacavir Example: Target Parameter and Observed Data

- Target causal parameter: $E_{U,X}(Y_{\bar{a}=1} - Y_{\bar{a}=0})$
- Observed data: n i.i.d. copies of
  $O = (L(1), A(1), L(2), A(2), Y) \sim P_0$
- Under what conditions can we write our target causal parameter as a parameter of the observed data distribution?
- **We need to move beyond the simple back door criterion….**

How are longitudinal parameters different?

- Our previous identifiability result relied on stratifying on some set of covariates W that were sufficient to block all back door paths from our intervention A to our outcome Y

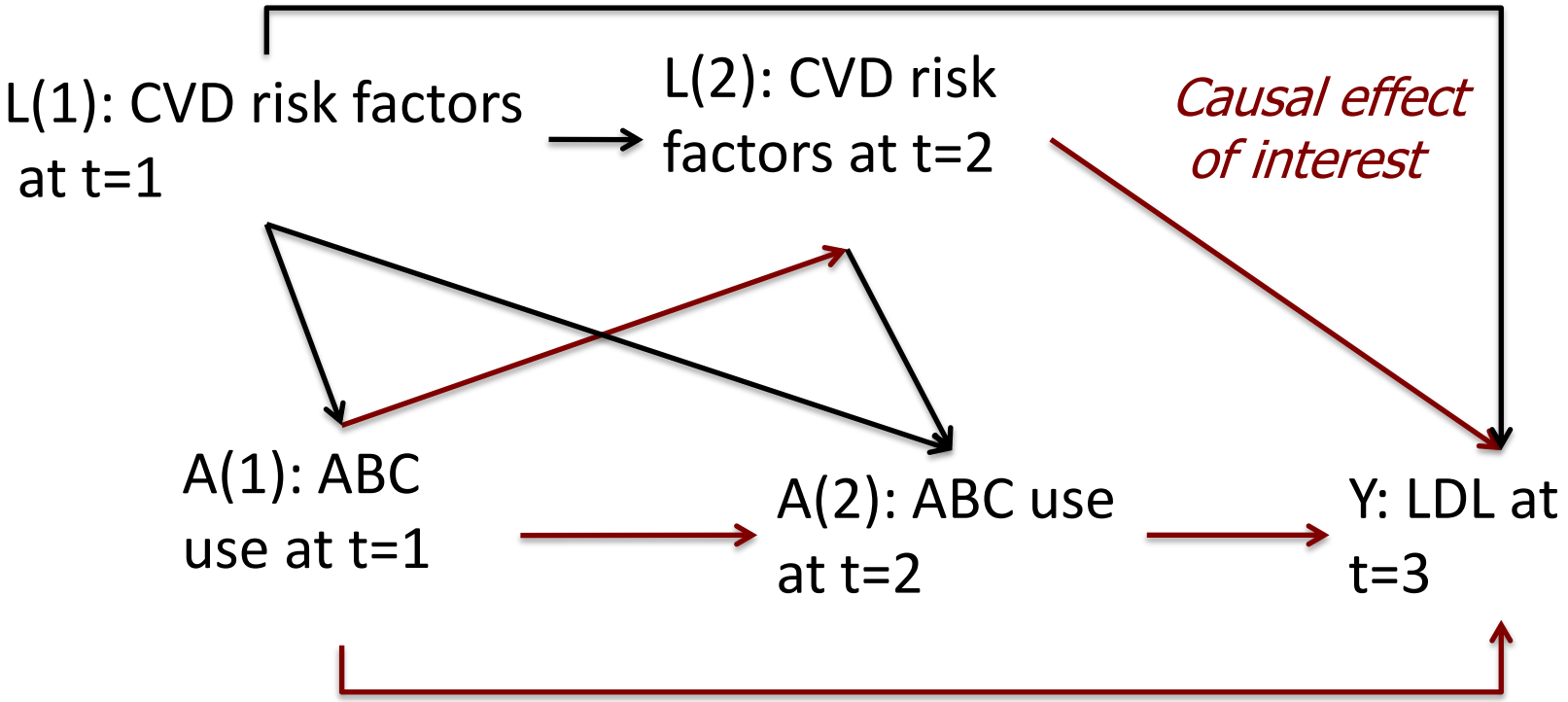
- We could not stratify on descendents of A

How are longitudinal parameters different?

- When we are interested in intervening on multiple nodes, we are often in a situation where no one set of covariates that meet the back door criterion for all intervention nodes simultaneously exists

- However, the distribution of counterfactuals indexed by interventions on these multiple nodes may still be identified…

ABC Example: SCM/Graph



$$L(t) = f_{L(t)}(\bar{L}(t-1), \bar{A}(t-1), U_{L(t)}), t = 1, 2, 3$$

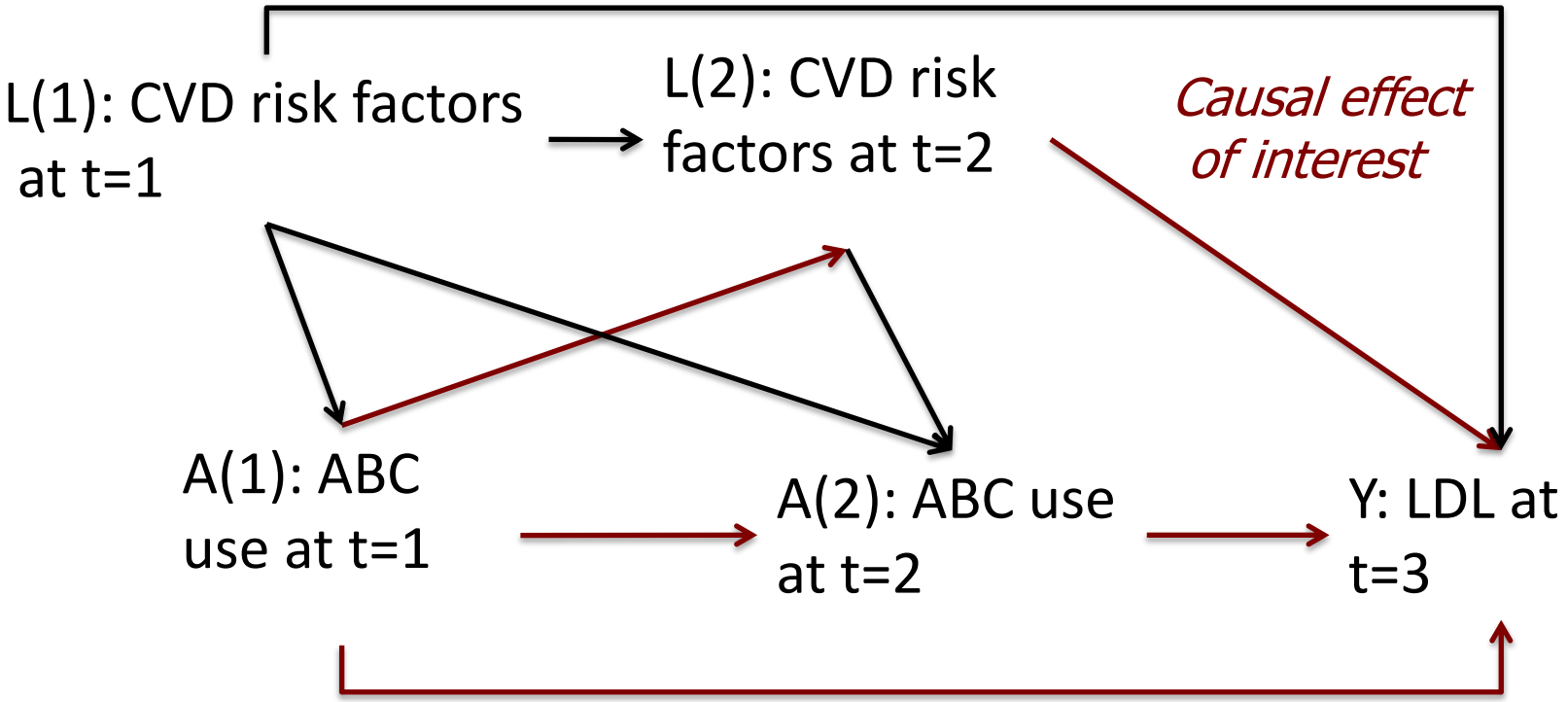
$$A(t) = f_{A(t)}(\bar{L}(t), \bar{A}(t-1), U_{A(t)}), t = 1, 2$$

$$Y \subset L(3)$$

Is $E(Y_{11} - Y_{00})$ identified using the standard (point treatment) back door criterion?

- We need to find a set of variables that
  1. Are non-descendents of (A(1),A(2)) and
  2. Block all back door paths from (A(1),A(2)) to Y…

ABC Example: SCM/Graph



$$L(t) = f_{L(t)}(\bar{L}(t-1), \bar{A}(t-1), U_{L(t)}), t = 1, 2, 3$$

$$A(t) = f_{A(t)}(\bar{L}(t), \bar{A}(t-1), U_{A(t)}), t = 1, 2$$

$$Y \subset L(3)$$

The Dilemma of Time –Dependent Confounding

- No subset of covariates for which the simple back door criterion holds
  - We need L(2) to block the back door path from A(2) to Y
  - But L(2) is a descendent of A(1)
- Is our target parameter really unidentified?

- Not necessarily! But we do need a new identifiability assumption -> new estimand

Key insight: we don't need to adjust for everything all at once

- Instead, we can think of simulating our data sequentially from our set of structural equations
- This lets us consider the problem of identifiability sequentially
  - For each A(t) in sequence, ask if its effect on Y can be identified by conditioning on some subset of the observed past.

Sequential Back Door Criterion

- Essentially we just want to apply the usual back door criterion, for each intervention node A(t) in series:
  1. We are looking for set of covariates (+ past treatment) that will block all back door paths from A(t) to the outcome
  2. These covariates cannot be descendents of A(t)
- Same justification: Want to remove any sources of association between each A(t) and the outcome other than those that we are interested in

Sequential Back door Criterion

- Just the standard back door criterion applied to each intervention node is sequence except
- Now it is OK if there is an unblocked back door path that goes through a future intervention node
- Why?
  - Any paths through future A nodes will already be blocked because we are intervening on them
  - We don't need to worry about blocking them

Identifiability for the effects of multiple interventions

- Intuition: Sequentially Randomized Trial
  - At each time point, randomize A(t) within strata of (some subset of) covariates and treatment observed up until then
  - In this case, at each time point the effect of A(t) on future nodes is identified
    - We know we measured enough of the past the estimate the effect of intervening on that node
  - We can estimate the effect of setting each A(t) sequentially

Identifiability for multiple interventions

- Sequential Randomization Assumption

$$Y_{\bar{a}} \perp A(t) | \bar{L}(t), \bar{A}(t-1) = \bar{a}(t-1), t = 1, ..., K$$

  - If A(t) is randomly assigned at each time point, given the observed past, this will hold
  - This is called a sequentially randomized trial or sequential multiple assignment randomized trial (SMART)

Identifiability Result

- Under Sequential Randomization Assumption, have the longitudinal G-computation formula:

$\Psi^F(P_{X,U,0})$: Target causal quantity

$$P(Y_{\bar{a}} = y) =$$

$$\sum_{\bar{l}} \left(\begin{array}{l} P(Y = y|\bar{A} = \bar{a}, \bar{L} = \bar{l}) \\ \prod_{t=0}^{K} P(L(t) = l(t)|\bar{A}(t-1) = \bar{a}(t-1), \bar{L}(t-1) = \bar{l}(t-1)) \end{array} \right)$$

$\Psi(P_0)$: Target statistical parameter/estimand

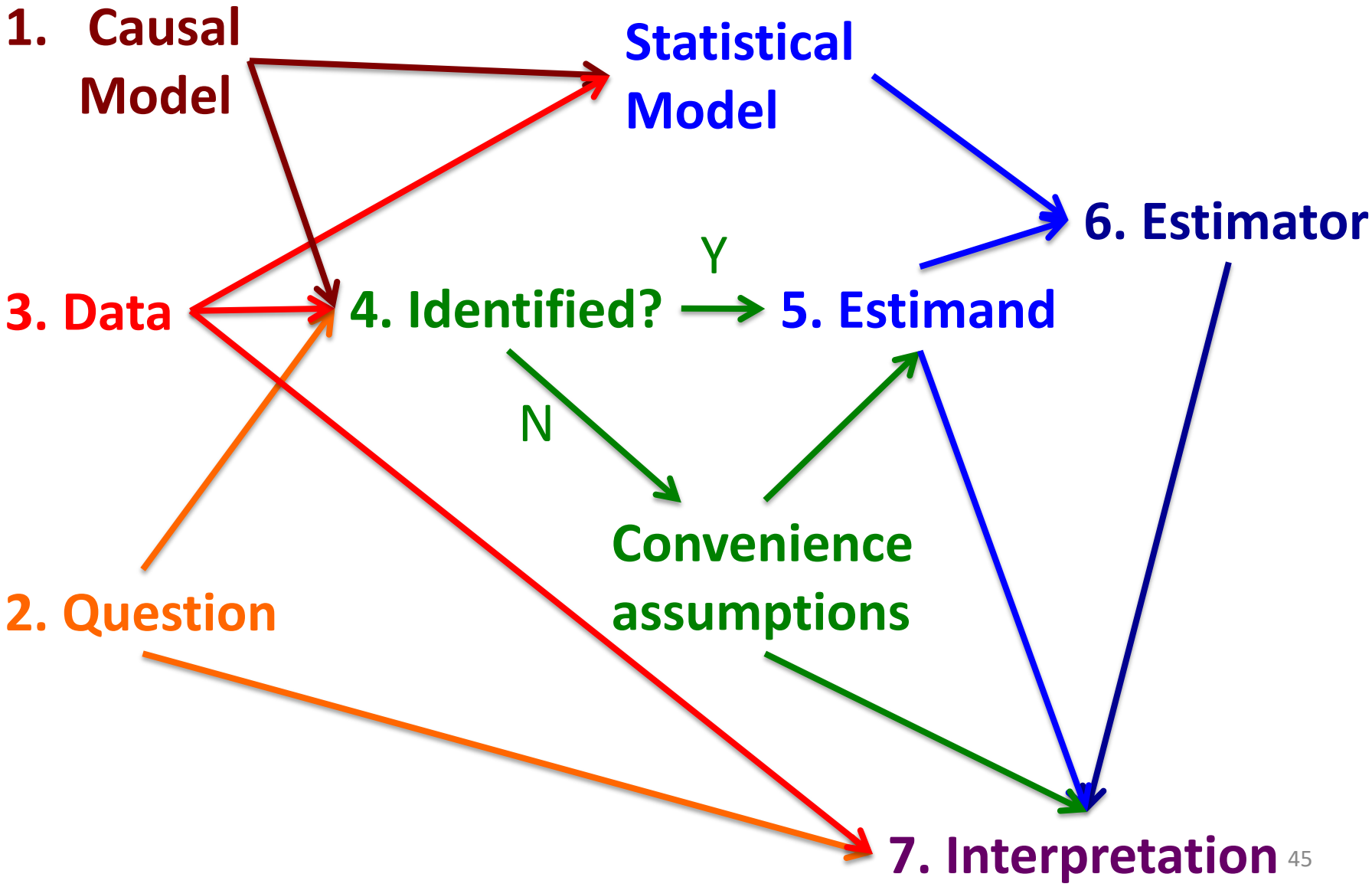
Positivity Assumption

- In order for $\Psi(P_0)$ to be defined (in a non-parametric model), need each treatment history of interest to occur with some positive probability for each possible covariate history

$$\min_{a \in \mathbf{A}} g(a(t) \mid \bar{A}(t-1), \bar{L}(t)) > 0 \ a.e.$$

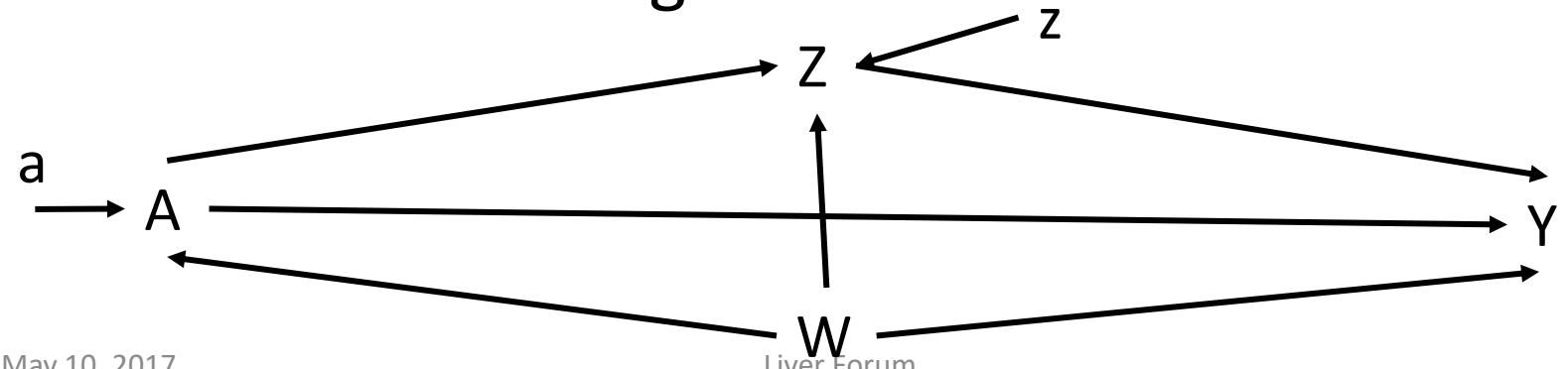
- Positivity violations are common
  - Some types of patients may develop absolute indications or contraindications for some treatments
    - Ex. g(ABC(t)=1|Contraindication(t))=0
  - Can also have lack of support in finite samples due to chance

A Roadmap….


1. Causal Model
Statistical Model
6. Estimator
Y
3. Data
4. Identified?
5. Estimand
N
Convenience assumptions
2. Question
7. Interpretation

Additional target causal quantities: Effect Mediation

- Interventions on more than one node can also be used to study effect mediation
- Single time point example:
  - How much of the effect of abacavir (A) on MI risk (Y) is due to changes in an inflammatory biomarker (Z)?
  - Define counterfactual outcome setting the levels of both treatment (A) and intermediate (Z): Y_{az}
- Generalizes to longitudinal data

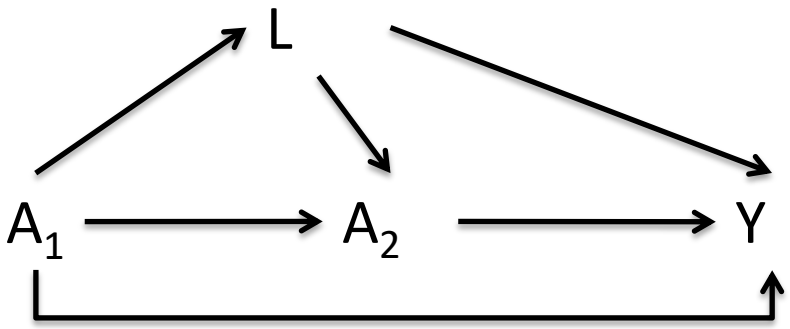


Effect Mediation

- Controlled Direct Effect: $E(Y_{1z}-Y_{0z})$
  - By fixing level of intermediate, effect of treatment on outcome cannot be mediated via changes in intermediate
  - Definition, identification and estimation results follow directly from those for longitudinal exposures (Robins 1999)
- Other effect mediation parameters involve nested counterfactuals
  - Z_a: counterfactual value of intermediate under treatment level a
  - Natural Direct Effect: $E(Y_{1Z_0} - Y_{0Z_0})$
  - Indirect Effect: $E(Y_{1Z_1} - Y_{1Z_0})$

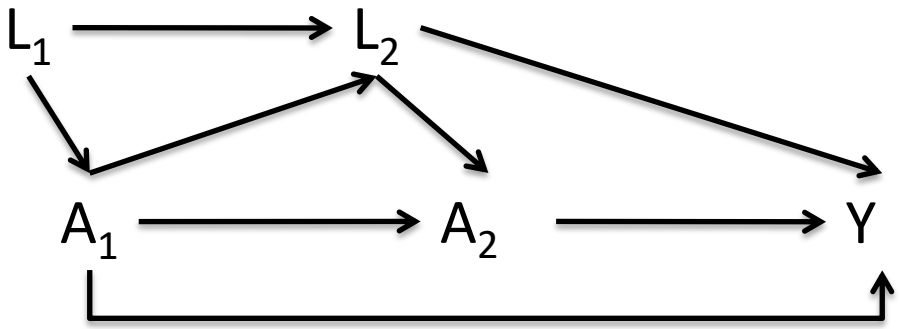
Example

- Target: $E(Y_{a1a2})$
- Sequential back door holds?
  - For A_1 given what?
  - For A_2 given what?



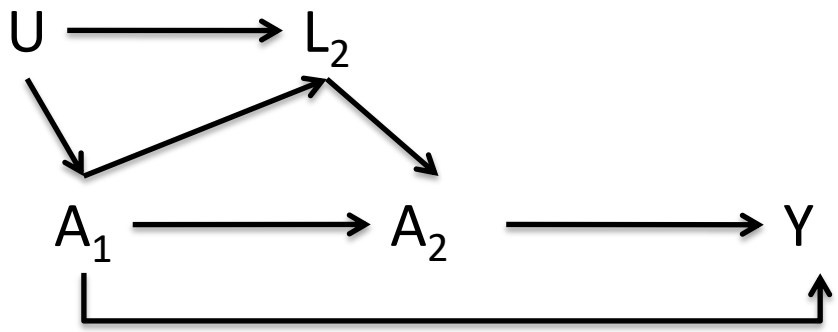
Example

- Target: $E(Y_{a1a2})$
- Sequential back door holds?
  - For A_1 given what?
  - For A_2 given what?



Example

- Target: $E(Y_{a1a2})$
- Sequential back door holds?
  - For A_1 given what?
  - For A_2 given what?



Example

- Target: $E(Y_{a1a2})$
- Sequential back door holds?
  - For A_1 given what?
  - For A_2 given what?