Lecture 2: Specifying a causal model
Structural Equations <-> Graphs

Outline

- Structural Causal Models (SCM)
- Non-parametric structural equation models
- Causal Graphs

References

- Petersen and van der Laan, Epidemiology 2014
- Pearl. "An Introduction to Causal Inference" *Int J Biostat*, 6(2): Article 7, 2010.
- Pearl. "The Causal Foundations of Structural Equation Modeling" in R. Hoyle, editor, Handbook of Structural Modeling. Guilford Press, New York, 2012
  - Associated technical report available at *ftp.cs.ucla.edu/pub/stat_ser/r370.pdf*

A roadmap for causal inference

1. Specify **Causal Model** representing real background knowledge
2. Specify **Causal Question**
3. Specify **Observed Data** and link to causal model
4. **Identify** : Knowledge + data sufficient?
5. Commit to an **estimand** as close to question as possible, and a **statistical model** representing real knowledge.
6. **Estimate**
7. **Interpret** Results

Definition: Structural Causal Model

1. Endogenous variables $X = \{X_1, ..., X_J\}$

– Variables that are meaningful for the scientific question, or about which you have some scientific knowledge

- E.g. We often (but not always!) know the time ordering of these variables
- Includes all the variables you measure (or are considering measuring)
- Might also include some variables you do not/cannot observe

– Affected by other variables in the model

Definition: Structural Causal Model

2. Background (exogenous) variables

$$U = \{U_1, ..., U_J\}$$

– Not affected by other factors in the model

– All the unmeasured factors not included in X that go into determining the values that the X variables take

  • U collapses all these unknown factors into one variable

– We denote the distribution of these factors P_U

Definition: Structural Causal Model

3. Functions $F = \{f_{X_1}, ..., f_{X_J}\}$

– The functions *F* define a set of structural equations for each of the endogenous variables

– For each endogenous variable in X_j, we specify its parents $Pa(X_j)$

- Which endogenous variables may affect what value X_j takes

$$X_j = f_{X_j}(Pa(X_j), U_{X_j}), j = 1, ..., J$$

$$Pa(X_j) \subseteq X \setminus X_j$$

Structural Causal Model

- Given an input *U*, the functions *F* deterministically assign a value to each of the endogenous variables

- Our model says that the distribution of *(U,X)* is generated by

1. Drawing a multivariate U from a specific probability distribution P_U
2. Deterministically assigning *X* by plugging *U* into the set of functions *F*

– A given input *u* gives us a specific realization *x*

Example: Effect of hormone replacement therapy (HRT) on cardiovascular disease (CVD)

- X={W, A, Y}
  - W=CVD Risk Factors,…
  - A=HRT use
  - Y= CVD
- Errors: $U=(U_W,U_A,U_Y) \sim P_U$
- Structural equations:

$$
\begin{aligned}
W &= f_W(U_W) \\
A &= f_A(W, U_A) \\
Y &= f_Y(W, A, U_Y)
\end{aligned}
$$

- <u>Distribution of (U,X) generated by:</u>

1. Draw *U* from P_U
2. Generate *W* as a deterministic function of U_W
3. Generate *A* as a deterministic function of *W* and U_A
4. Generate *Y* as a deterministic function of *W*, *A*, U_Y

What have we accomplished so far?

- We have defined a system for deterministically assigning values to a set of variables based on a random input

- Why is this useful or interesting?

SCM encode causal assumptions

- Assumptions about how the variables X were generated in the system we want to study
- What factors does "Nature" (or the "experiment" that generated the data in the system we want to study) consult when assigning a value to these variables?
  - What factors affect physicians' and patients' decisions to start HRT?
  - What are major determinants of cardiovascular outcomes in this population?

Structural Equations are not your everyday equations!

- Regular equations are symmetric
  - $X_2=X_1+U \rightarrow X_1=X_2-U$
- Causality (and structural equations) are not
- Example:
  - Assume: Headache=Brain tumor + U
    - U=(Flu, red wine, genetic predisposition to migraines, stress…)
  - Does not imply: Brain tumor=Headache – U
    - Headaches do not cause brain tumors

SCM encode causal assumptions

- Assumption type #1: Exclusion restrictions

$$X_j = f_{X_j}(Pa(X_j), U_{X_j}), j = 1, ..., J$$
$$Pa(X_j) \subseteq X \setminus X_j$$

  - Restrictions on the Parents of X
    - Which variables appear in the right hand side of each structural equation
  - **Common point of confusion** "exclusion restriction" when applied to a causal model does not refer to criteria for excluding a subject from a study

More on exclusion restrictions

- We make assumptions by leaving X variables out of a given parent set
  - Excluding a variable from Pa(X_j) assumes it does not directly affect what value X_j takes
  - Leaving a variable in Pa(X_j) just means it might affect what value X_j takes

$$X_j = f_{X_j}(Pa(X_j), U_{X_j}), j = 1, ..., J$$
$$Pa(X_j) \subseteq X \setminus X_j$$

What does our model assume?

- Example 1:

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Y &= f_Y(W, U_Y) \end{aligned}$$

W=Flu virus
A= Headache
Y=Cough

- Example 2:

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(U_A) \\ Y &= f_Y(W, A, U_Y) \end{aligned}$$

W= Parental education
A= Random selection to receive school voucher
Y=Test scores

SCM encode causal assumptions

- Assumption type #2: Independence assumptions
  - Restrict the allowed distributions for P_U
- Ex. U_A is independent of U_Y
- Corresponds to saying that A and Y share no common causes outside of any in X
  - When might this be reasonable?

Assume U_A independent of U_Y?

- Example 1:

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Y &= f_Y(W, U_Y) \end{aligned}$$

W=Flu virus
A= Headache
Y=Cough

- Example 2:

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(U_A) \\ Y &= f_Y(W, A, U_Y) \end{aligned}$$

W= Parental education
A= Random selection to receive school voucher
Y=Test scores

More on independence assumptions

- U provide the random input into a deterministic system
- The sources of this randomness are generally not something we understand well
  - At this point we avoid making any unsupported assumptions on the joint distribution of the errors
  - Sometimes assumptions on U will be supported by knowledge…

Example: Effect of hormone replacement therapy (HRT) on cardiovascular disease (CVD)

- X={W, A, Y}
  - W=CVD Risk Factors,…
  - A=HRT use
  - Y= CVD
- Errors: $U=(U_W,U_A,U_Y) \sim P_U$
- Structural equations:

$$
\begin{aligned}
W &= f_W(U_W) \\
A &= f_A(W, U_A) \\
Y &= f_Y(W, A, U_Y)
\end{aligned}
$$

- <u>Distribution of (U,X) generated by:</u>

1. Draw *U* from P_U
2. Generate *W* as a deterministic function of U_W
3. Generate *A* as a deterministic function of *W* and U_A
4. Generate *Y* as a deterministic function of *W*, *A*, U_Y

What does it mean to "draw U"

- These unmeasured errors have a distribution in the underlying population
  - Ex. Joint distribution of health care access, personal preferences towards HRT use, income, etc.
- Drawing a subject from the population corresponds to drawing a particular realization *u* of these background factors *U*
  - The subject drawn has specific values for each factor in *U*, regardless of whether we measure them or even know what they are

What variables to include in X?

- What variables go into X (and thus get explicit structural equations) will depend on your knowledge, the system you want to study, and your question
- Ex: where does genetics go?
  - If you are interested in the effect of vitamins on birth outcomes?
  - If you are interested in maternal genetic factors that predispose towards poor birth outcomes?

What variables to include in X?

- We haven't formally introduced the observed data yet. But we should be thinking ahead. If there is a variable we know we have access to, worth adding it as an X.
  - Forces us to think explicitly about its role
  - It may be useful (or essential) for estimation
- That said, don't need to limit to observed variables.
  - Ex. Smoking is known to be an important determinant of CVD and HRT use

U vs. X

- U is the random input into our system
  - We will never observe all the elements in U
  - Always some sources of randomness we don't understand
- That doesn't mean it might not be possible to observe some of the elements in U
  - As we learn more, variables can move from U to X
  - Ex- we learn about genetic factors and start including them as X variables in our models (and maybe measuring them)

SCM also called: "Non-Parametric" Structural Equation Models

- The structural equations do not restrict the functional form of the causal relationships
  1. We have specified $Y=f_Y(A,W,U_Y)$
  2. We have not specified $Y=\beta_1 A+\beta_2 W+U_Y$
- Historically, most structural equation modeling has done the latter
  - What does equation 2. assume?
  - When might this be a problem…
    - A=HRT dose, W=age, Y=Blood pressure
- If you have such knowledge however, you can and should incorporate it!

A word about “Models”

- Statistical model: set of allowed distributions for the observed data
  - Simple Example: One Random variable- Age
  - Your model might assume it is normally distributed
    - This restricts its set of possible distributions to the set of normal distributions
  - Or your model might put no restrictions on its distribution….

Structural Causal Model

- Defines set of allowed distributions for (U,X)
- Specifically, this is the set of distributions defined by
  - All the joint distributions P_U compatible with our independence assumptions
  - All the specifications of the functions $F=(f_{Xj}: j)$ compatible with our exclusion restrictions
- We will call this model $\mathcal{M}^{\mathcal{F}}$
  - Each distribution included in the model is indexed by a specific distribution P_U and specific functions F

More on assumptions to come…

- Assumptions (at least on P_U) will be necessary if we want to make causal inferences with observational data- will come back to this when we talk about identifiability
  - Our goal- to keep these as minimal as possible, and understand what they are so we can evaluate them better
  - Any time you use, eg, multivariable regression to go after an causal effect you are making assumptions on your Us…
    - We are trying to make these explicit

From Structural Equations to a Causal Graph

Recap: Structural Causal Model

1. Endogenous variables $X = \{X_1, ..., X_J\}$
   - Affected by other variables in the model
   - May or may not be observed
2. Background (exogenous) variables $U = \{U_1, ..., U_J\}$
   - Not affected by other factors in the model
   - Not observed

– We denote the distribution of these P_U

3. Functions $F = \{f_{X_1}, ..., f_{X_J}\}$
   - The functions *F* define a set of structural equations for each of the endogenous variables

$$X_j = f_{X_j}(Pa(X_j), U_{X_j}), j = 1, ..., J$$

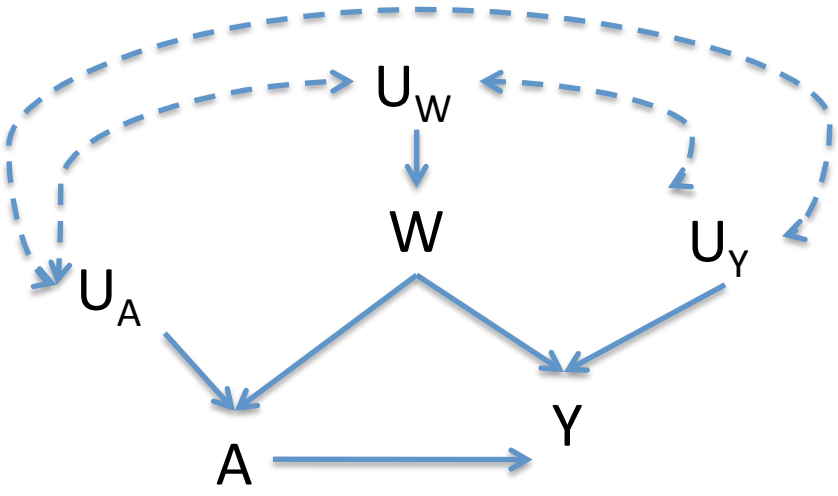
$$Pa(X_j) \subseteq X \setminus X_j$$

Graph Terminology

- Each Variable also referred to as a node
- Families
  - A is a parent of Y, Y is a child of A
  - A and Y are descendents of U_A
  - U_A is an ancestor of A and Y
- Edge= line or arrow connecting two nodes
- Path = any consecutive sequence of edges

Structural Model Defines a Graph

$$
\begin{aligned}
W &= f_W(U_W) \\
A &= f_A(W, U_A) \\
Y &= f_Y(W, A, U_Y)
\end{aligned}
$$



- Connect parents to children with an arrow
  - Makes the asymmetry of the equations explicit
- Each endogenous variable has an error (U)
- Potential dependence between errors encoded in dashed lines/double headed errors.

A note on feedback loops

- In this class we will work with acyclic systems
  - i.e. No directed cycles/feedback loops.
  - “recursive”
- SCM can also be non-recursive…
  - Dealing with acyclic systems simplifies things
  - Doesn’t have to be very limiting…
- We can incorporate feedback loops through temporal ordering

More formally…

- Before we gave a general SCM definition
- We will work with recursive SCM
  - Recursive Model: There exists an ordering

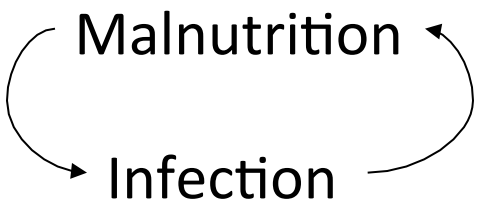
  $X=\{X_1,\ldots,X_J\}$ such that each X_j is a function of a subset $Pa(X_j)$ of its predecessors

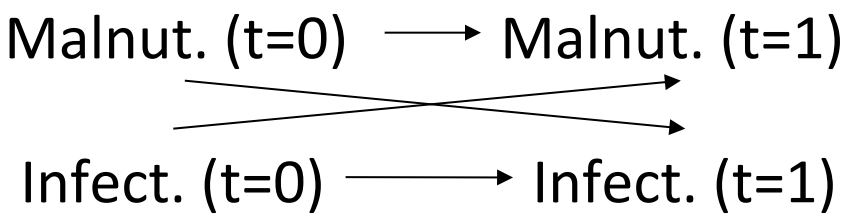
$$X_j = f_{X_j}(Pa(X_j), U_{X_j}), j = 1, ..., J$$
$$Pa(X_j) \subseteq \{X_1, ..., X_{j-1}\}$$

  - A natural source of ordering: time

Dealing with feedback…

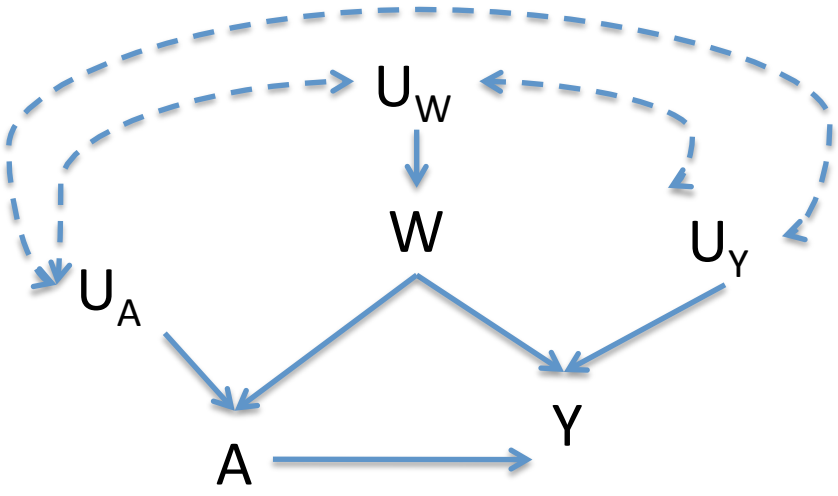
- Causes always precede their effects
  - This is true regardless of whether we observe this time ordering…
- To avoid feedback loops, extend graph (and corresponding structural equations) over time





Exclusion restrictions on the graph

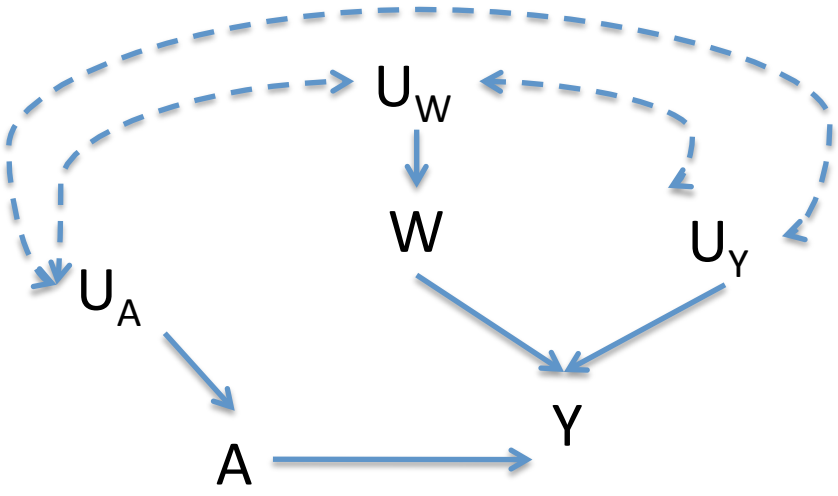
$$
\begin{aligned}
W &= f_W(U_W) \\
A &= f_A(W, U_A) \\
Y &= f_Y(W, A, U_Y)
\end{aligned}
$$



- Encoded through absence of arrows between X variables
  - Specification of parents in structural equation
  - Absence of arrow means no direct effect

Exclusion restrictions on the graph

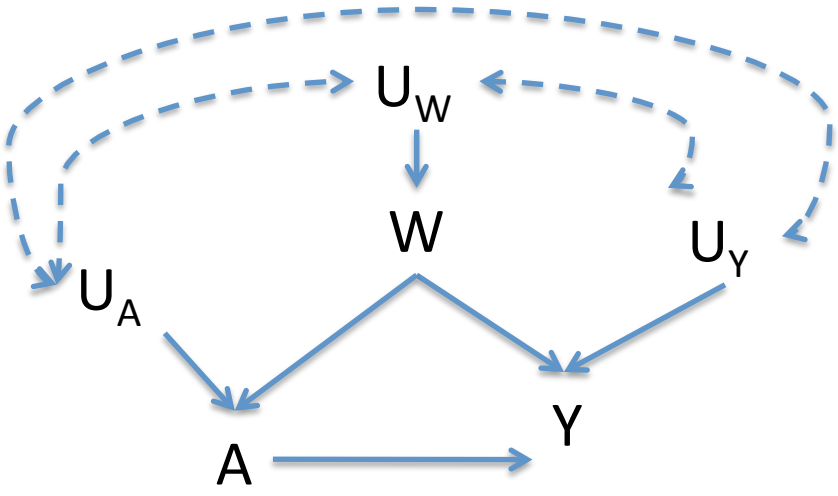
$$W = f_W(U_W)$$
$$A = f_A(U_A)$$
$$Y = f_Y(W, A, U_Y)$$



- Encoded through absence of arrows between X variables
  - Specification of parents in structural equation
  - Absence of arrow means no direct effect

Independence assumptions on the graph

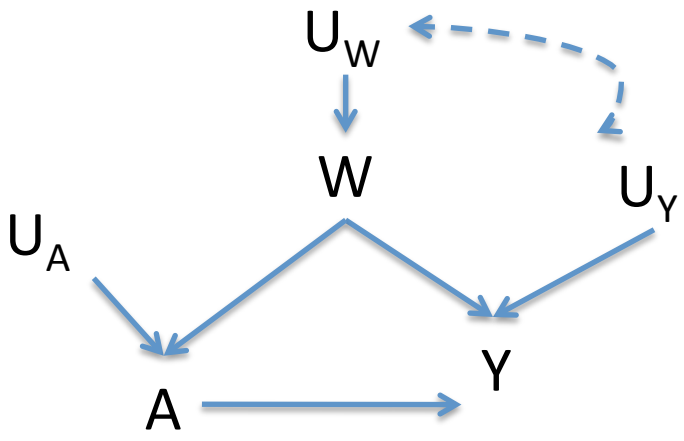
$$
\begin{aligned}
W &= f_W(U_W) \\
A &= f_A(W, U_A) \\
Y &= f_Y(W, A, U_Y)
\end{aligned}
$$



- Absence of double headed arrows between U
  - Assumption on distribution P_U
- Absence of double headed arrow means those two errors independent
  - No unmeasured shared common cause

Independence assumptions on the graph

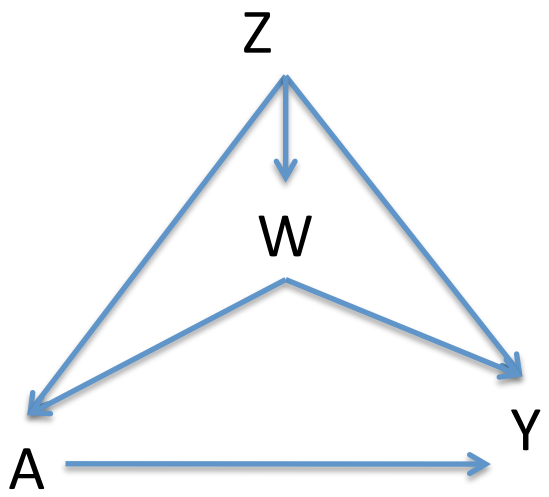
$$
\begin{aligned}
W &= f_W(U_W) \\
A &= f_A(W, U_A) \\
Y &= f_Y(W, A, U_Y)
\end{aligned}
$$



- Absence of double headed arrows between U
  - Assumption on distribution P_U
- Absence of double headed arrow means those two errors independent
  - No unmeasured shared common cause

Alternative representation: Independence assumptions

$$Z = f_Z(U_Z)$$
$$W = f_W(Z, U_W)$$
$$A = f_A(Z, W, U_A)$$
$$Y = f_Y(Z, A, W, U_Y)$$



- Include as a node any unmeasured common cause of at least 2 of the X variables
  - Doesn't have to represent a specific variable that you understand well
  - Just an alternative way to express that there may be such a variable (or variables)
- The remaining errors will be independent
  - Customarily omitted from the graph
- Resulting graph is "directed" as well as "acyclic" (DAG)

Graphs vs. Structural Equations

- Graph is an excellent tool for communicating with subject matter experts
  - Without successful communication, your analysis will not be worth much
- Graph can be a helpful way to translate assumptions into a formal model
  - As we will see, however, as the model becomes more complicated, the equations get a lot more friendly to work with

Graphs vs. Structural Equations

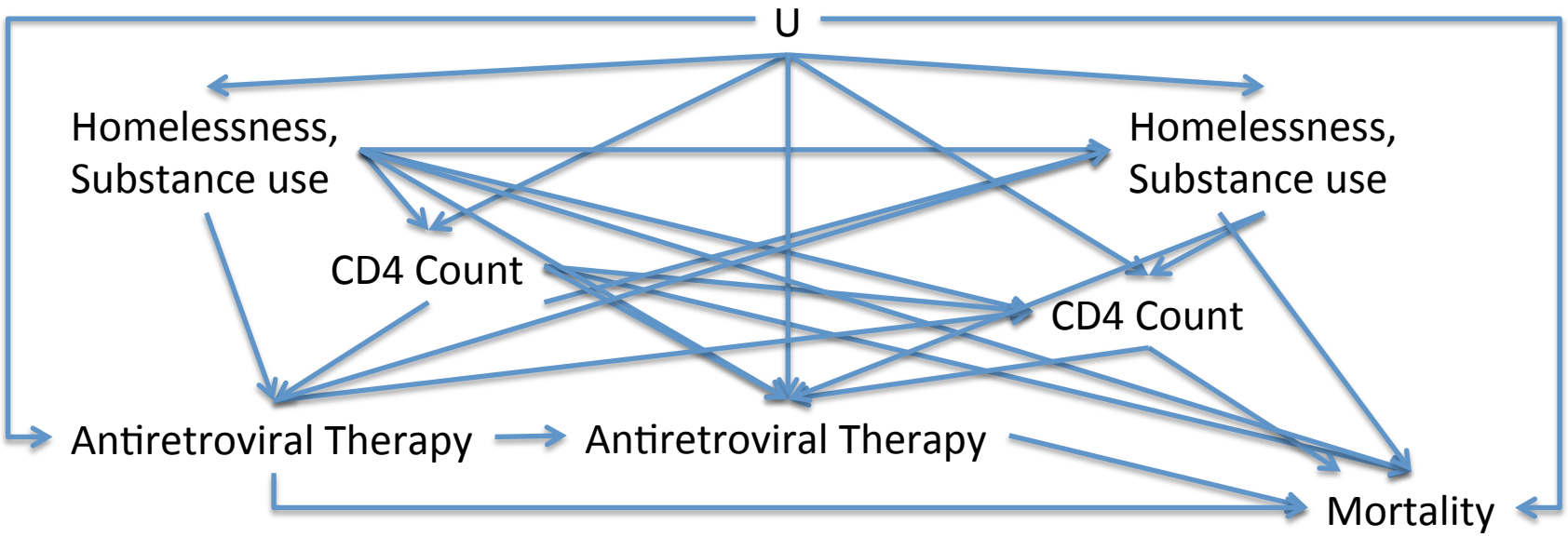
- Graphs are also be very useful for reading off what restrictions (if any) our model puts on the joint distribution of the observed data
  - These are testable assumptions
  - This tells us what statistical model is implied by the causal model
  - Tool for assessing identifiability, suggesting estimation approaches

Graphs vs. Structural Equations

- Equations can be easier to work with
  - Especially when you get into longitudinal data
- Equations are great for honing your identifiability assumptions
  - Figuring out the minimal you need to assume to move from a parameter of the distribution of (U,X) to a parameter of the distribution of the observed data
- Equations may help you resist the urge to oversimplify…

All causal models are not wrong…

- Causal models should represent real knowledge
  - We often need to make additional assumptions in order to make progress
  - Keep this process separate



What have we accomplished so far?

- We have formally defined a model that expresses our assumptions about a system we want to study

- Why is this useful or interesting?

- Our goal is to move from a real world question to a parameter of the distribution of the observed data…

Some orientation

- We have not defined our observed data yet.
- We have specified a causal model encoding our knowledge about a system out there in the world
- Next we will formally express what we would like to learn about this system.
- After that we will talk about assessing the feasibility of learning what we would like to learn (identifiability), what else we might need to assume or measure to make it feasible, and how to go about it (estimation).

Key Take Home Points

1. Causal Inference requires background knowledge
2. Causal Models represent background knowledge formally
3. Structural Causal Model
   - Represented as a set of non-parametric structural equations
   - or as a graph
   - Represents what we know AND what we do not know
   - Knowledge represented by
     1. Exclusion restrictions (missing arrows)
     2. Independence assumptions (independence of background factors, or no unmeasured common cause)

Coming up next…

- Translating scientific questions into formal target causal parameters
- Causal effects defined using interventions on the SCM
- The link between SCM and counterfactuals
  - Unifying two approaches to causal inference and getting the most out of each
- Marginal Structural Models to define target causal parameters