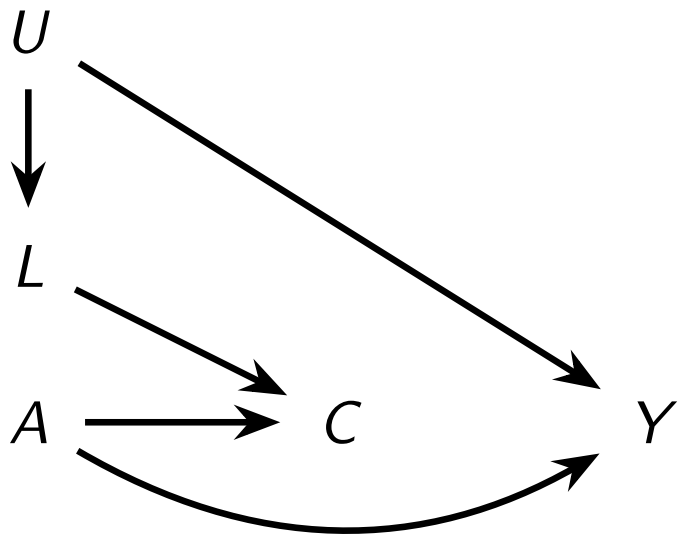
Section 20

Lecture 6

Loss to follow-up example 1



Factorisation according to the DAG with ordering $\langle A, U, L, C, Y \rangle$:

$$p(y, c, l, u, a) = p(y \mid u, a)p(c \mid l, a)p(l \mid u)p(u)p(a)$$

But how do we use this factorization to identify causal effects?

A clinical story

- Suppose the graph on Slide 163 represents a study of HIV-positive individuals to estimate the effect of an antiretroviral treatment A on 3-year risk of death Y.
- The unmeasured variable $U \in \{0, 1\}$ indicates high level of immunosuppression. Those with $U = 1$ have a greater risk of death.
- Individuals who drop out from the study or are otherwise lost to follow-up are censored ($C = 1$).
- Individuals with $U = 1$ are more likely to be censored because the severity of their disease prevents them from participating in the study.
- The effect of U on censoring C is mediated by the presence of symptoms (fever, weight loss, diarrhea, and so on), CD4 count, and viral load in plasma, all included in L, which we suppose are measured.
- Individuals receiving treatment are at a greater risk of experiencing side effects, which could lead them to dropout, as represented by the arrow from A to C. We have to restrict the analysis to individuals who remained uncensored ($C = 0$) because those are the only ones in which Y can be assessed.

Backdoor theorem in the example on loss to follow-up

Consider the example from Slide 163.

- Note that
  - L blocks all backdoor paths between (A, C) and Y.
  - Thus,

$$\mathbb{E}(Y^{a,c=0}) = \sum_l \mathbb{E}(Y \mid A = a, C = 0, L = l)P(L = l),$$

which can be estimated simply by standardisation:

- Estimate $\mathbb{E}(Y \mid A = a, C = 0, L = l)$ by $\hat{\mathbb{E}}(Y \mid A = a, C = 0, L = l)$,
- Estimate $P(L = l)$ empirically.

The standardisation estimator is:
$\frac{1}{n}\hat{\mathbb{E}}(Y^{a,c=0}) = \sum_i \hat{\mathbb{E}}(Y \mid A = a, C = 0, L = L_i)$

PS: Many causal questions are more difficult

Realistic questions are often more difficult. Consider for example:

- when should we start a treatment?
- How long should we continue treatment?
- When to switch to different treatment?
- What event should guide us to switch treatment?

PS: Many causal questions are more difficult

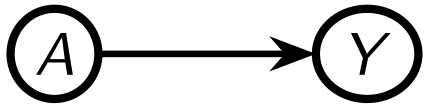
Realistic questions are often more difficult. Consider for example:

- when should we start a treatment?
- How long should we continue treatment?
- When to switch to different treatment?
- What event should guide us to switch treatment?

We will discuss such questions later in the course

Elephant in the room...

In a randomised study, the following graph is a causal DAG:



And we know that $Y^a \perp\!\!\!\perp A$ for $a \in \{0, 1\}$.
But the counterfactual independence cannot be read off from the graph!
This raises some questions:

- Can we construct graphs to read off such counterfactual independencies?
- Can we read off factorisations of *counterfactual* laws from graphs?

D-separation allows us to read off whether an association is causal

- We can graphically check – using d-separation – whether an observed association between two variables A and B conditional on C is (solely) due to a causal effect (i.e. that the association is unconfounded).
- However, we also want to use graph to evaluate if we can identify functionals of *counterfactual* variables, for example $\mathbb{E}(Y^a)$. We can use the backdoor theorem for this task, but the elephant in the room is that there are no counterfactual variables on the DAG! And we did want to reason about counterfactual independencies. Thus, whereas we can evaluate independencies between *factual* variables in a DAG, we cannot study *counterfactual* independencies.
- Here we will study a recent and elegant[23] transformation of the DAG – the so-called Single World Intervention Graph (SWIG) – that does allow us to read off independencies between factual and counterfactual variables.

[23] Thomas S Richardson and James M Robins. "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality". In: *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper* 128.30 (2013).

Section 21

Single World Intervention Graphs (SWIGs)

Creation of SWIGs

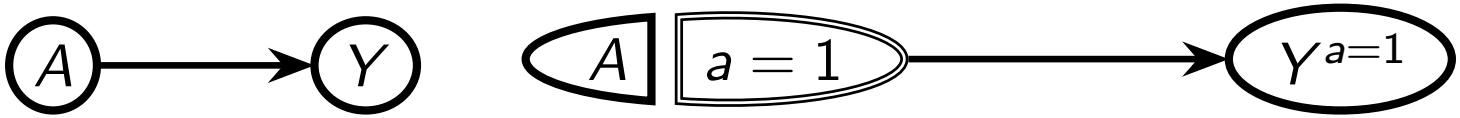
The SWIG $\mathcal{G}(a)$ is created as follows:

1. Draw the DAG that represents the causal model.
2. Split treatment variables into two nodes (indicated by semi-circles), left and right.
   - The left node encodes the random variable treatment that would have been observed in the absence of an intervention. This is called the *natural value of treatment* node. Natural value of treatment nodes should be treated as nodes of an ordinary DAG, i.e., ordinary random variables.
   - The right node encodes the value of treatment under the intervention. These nodes should be treated as constants, i.e. fixed nodes.
3. Re-label every non-manipulated descendant of an intervention node with superscript: the superscripts indicate the counterfactual.
   - Use consistency to obtain graphs with minimal labelling, i.e. the minimal set of counterfactuals in the superscript.

The SWIG can be conceived as a function that transforms the original causal DAG into a new graph, which is still (formally) a DAG.

Example: SWIG in a simple randomised trial

SWIG under treatment $a = 1$:



We can read the independence $Y^{a=1} \perp\!\!\!\perp A$.
We also associate the new **factorisation**:

$$P(A = a', Y^{a=1} = y) = P(A = a')P(Y^{a=1} = y),$$

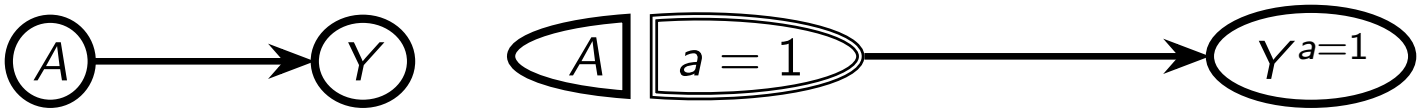
where we omit the fixed nodes from the conditioning set. Furthermore, we make a **modularity** assumption

$$P(Y^{a=1} = y) = P(Y = y \mid A = 0),$$

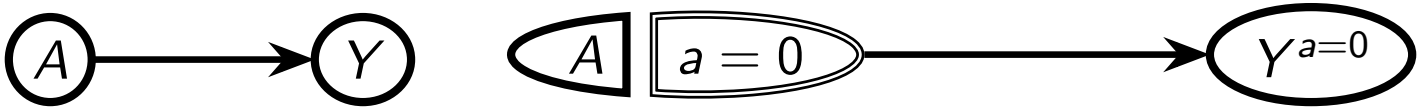
which links the original factorisation to the original DAG factorisation. This modularity assumption is indeed implied by the consistency assumption, which is in turn implied by independent error assumption in the NPSEM-IE.

Single world

We can read the independence $Y^{a=1} \perp\!\!\!\perp A$ from the SWIG for treatment $a = 1$:



We can read the independence $Y^{a=0} \perp\!\!\!\perp A$ from the SWIG for treatment $a = 0$:



Why do we need both graphs? These are two different graphs that represent the factorisation of different margins: $P(A = a', Y^{a=1} = y)$ and $P(A = a', Y^{a=0} = y)$. None of the SWIGs encodes assumptions between counterfactuals from different worlds $Y^{a=0}$ and $Y^{a=1}$. This is a feature, not a bug.
It has to do with identification. Node splitting preserves identification. If I observe every node that I included in the original DAG, then the counterfactual laws defined by the node splittings are also going to be identified.
For example, if in the DAG above $P(A = a', Y = y)$ is identified, then $P(A = a', Y^{a=1} = y)$ is identified and so is $P(A = a', Y^{a=0} = y)$, but not $P(A = a', Y^{a=1} = y', Y^{a=0} = y)$.

Factorisation

Definition (SWIG factorisation)

The factorisation associated with a SWIG is

$$P(V^{\overline{a}} = v) = \prod_{V_i \in V} P(V_i^{\overline{a}_i} = v_i \mid (PA_{\mathcal{G}(\overline{a}),i} \setminus \overline{a}) = q)$$

where $q \subseteq pa_i \subset v$ and $\overline{a}_i \subseteq \overline{a}$ ($\overline{a}_i$ are the elements of $\overline{a}$ that are ancestors of V_i).

Modularity

Definition (Modularity)

The DAG pair $(\mathcal{G}, p(v))$ and the SWIG pair $(\mathcal{G}(\overline{a}), p^{\overline{a}}(v))$ under an intervention that sets $\overline{A} = (A_0, \ldots, A_k)$ to $\overline{a} = (a_0, \ldots, a_k)$ satisfy modularity for every $V_i \in V$ if

$$
\begin{aligned}
&P(V_i^{\overline{a}_i} = v_i \mid (PA_{\mathcal{G}(\overline{a}),i} \setminus \overline{a}) = q) \\
=&P(V_i = v_i \mid (PA_{\mathcal{G},i} \setminus \overline{A}) = q, (PA_{\mathcal{G},i} \cap \overline{A}) = \overline{a}_{PA_{\mathcal{G},i} \cap \overline{A}})
\end{aligned}
$$

This definition looks like a mouthful, but it is conceptually quite easy to understand. It bridges counterfactual densities to observable densities.
It is implied by the independent error assumption of the NPSEM-IE, and it holds under a weaker causal model, the FFRCISTG[24] (I have not shown this).

[24] Richardson and Robins, "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality".

Causal models, factorisation and modularity

Theorem

A NPSEM-IE model (and the FFRCISTG model that includes the NPSEM-IE model as a strict submodel) obeys factorisation and modularity.

We will not prove this result, but we will use it extensively.
In our saturated graph when we intervene to set $a = 1$, it implies that $P(Y^{a=1} = y) = P(Y = y \mid A = 1)$.

D separation of a path in a SWIG)

This definition is **very** similar to the definition in DAGs:

Definition (d-separation of a path)

A path r is d-separated by a set of nodes Z iff

1. r contains a chain $V_i \to V_j \to V_k$ or a fork $V_i \leftarrow V_j \to V_k$ such that V_j is in Z, or
2. r contains a collider $V_i \to V_j \leftarrow V_k$ such that V_j *is not* in Z and such that no descendant of V_j is in Z.

If a path is not d-separated by Z **and** there is no fixed node on the path, then the path is d-connected given Z.

SWIT in a simple randomised trial (experiment)

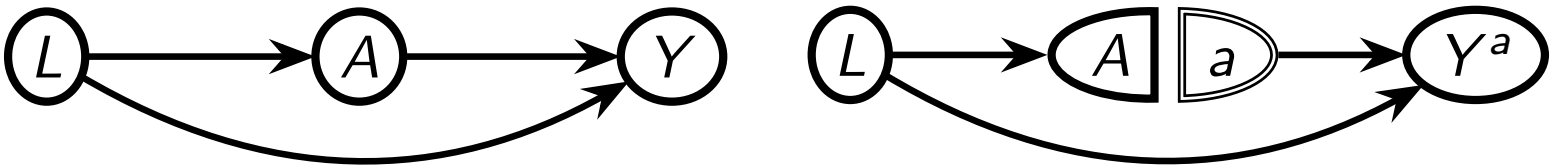
A SWIT is a SWIG template[25], i.e. a graph valued function:

- It takes a specific value a as input.
- Returns a SWIG $G(a)$.
- SWIG $G(0)$ represents $p(A = a', Y^{a=0} = y)$.
- SWIG $G(1)$ represents $p(A = a', Y^{a=1} = y)$.

The SWIT represents both the SWIGs from the previous slide. Hereafter we will use SWITs for simplicity, most of the time.

[25] Note that I am sometimes sloppy and use the word SWIG when I formally talk about a SWIT.

SWIG in a conditional randomised experiment

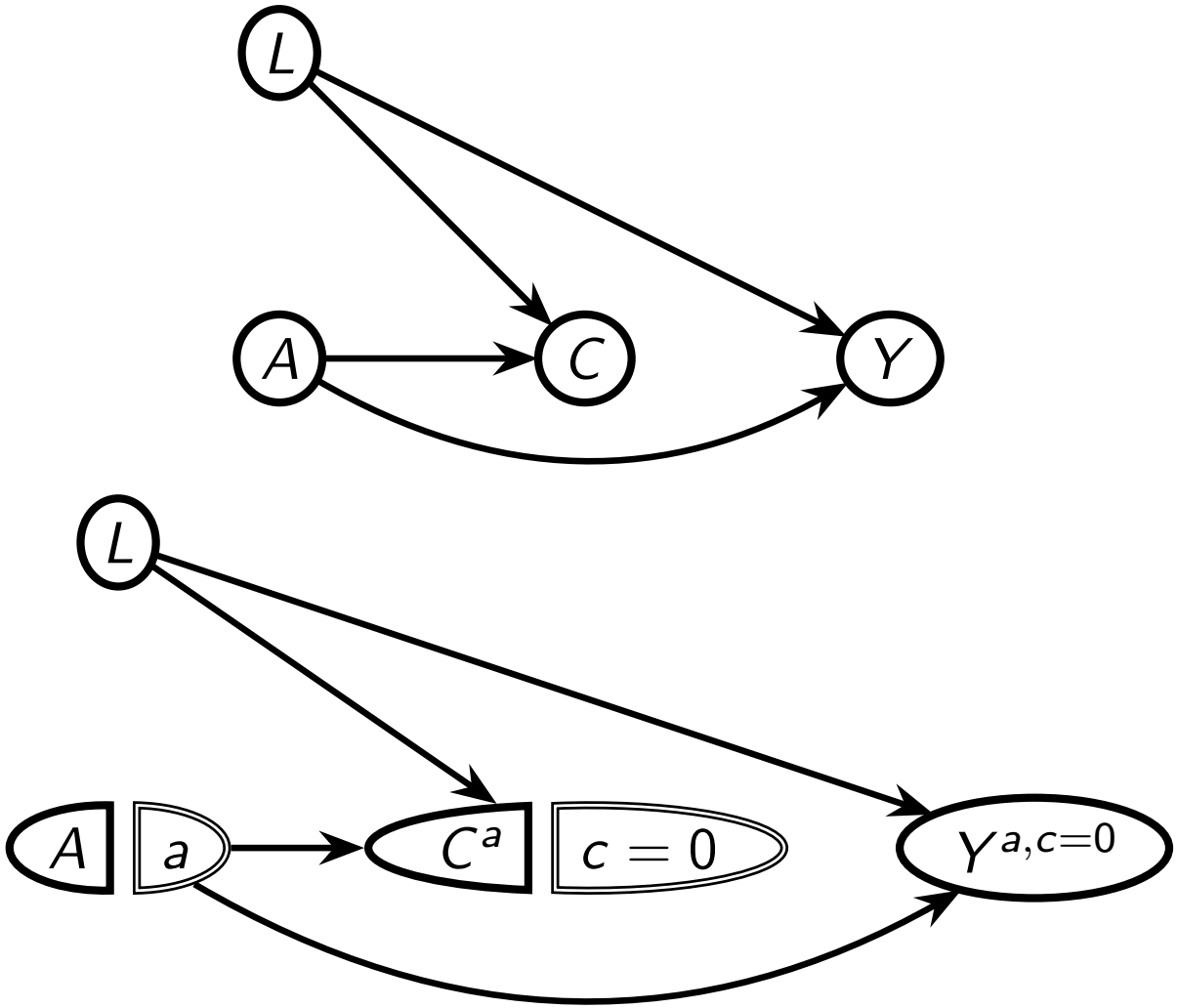


$$P(Y^a = y) = \sum_l P(Y^a = y \mid L = l)P(L = l) \text{ factorization}$$

$$= \sum_l P(Y = y \mid A = a, L = l)P(L = l). \text{ modularity}$$

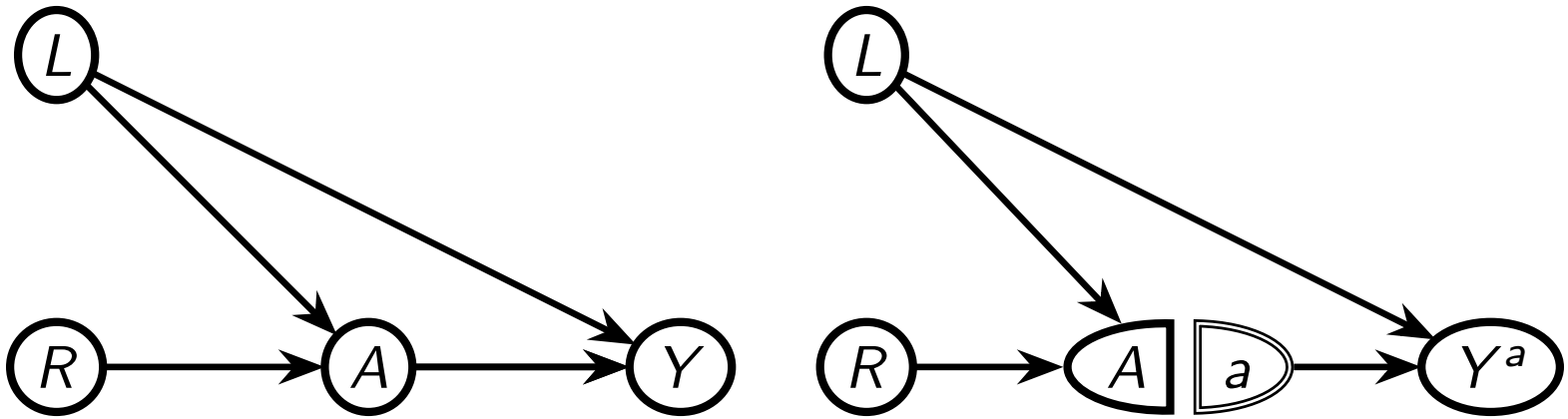
SWIG in an experiment with loss to follow-up (C)

A is treatment, C is censoring. The counterfactual outcome $Y^{a,c=0}$ is the outcome if we kept every individual uncensored ($c = 0$) under treatment a.



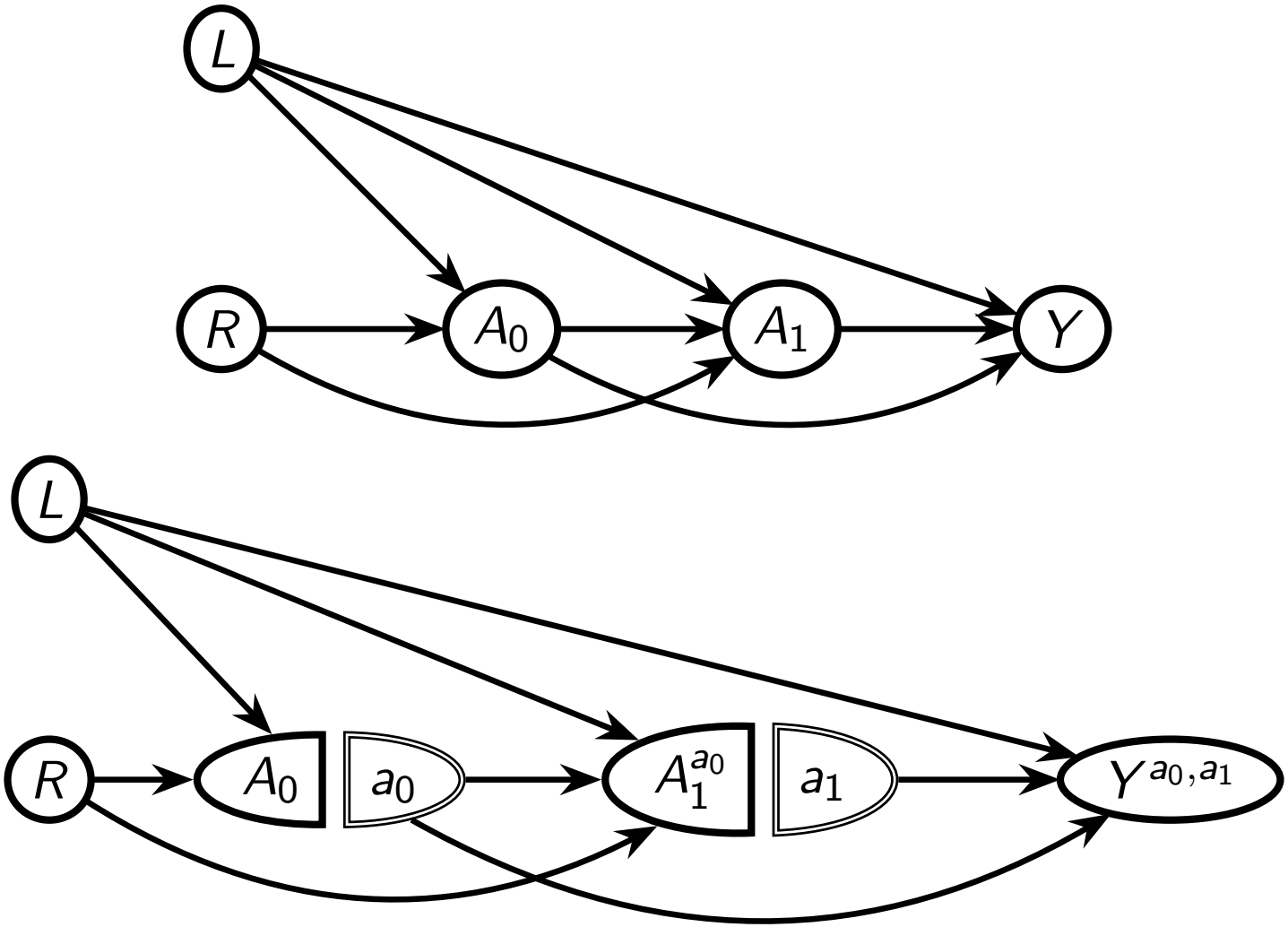
SWIG in an experiment with imperfect adherence

R is the strategy that was assigned, and A denotes taking treatment. Here, the counterfactual in the SWIG is the outcome had the patient taken treatment a. The lack of an arrow from R to Y^a encodes the assumption that randomisation only causes the outcome through the treatment A.



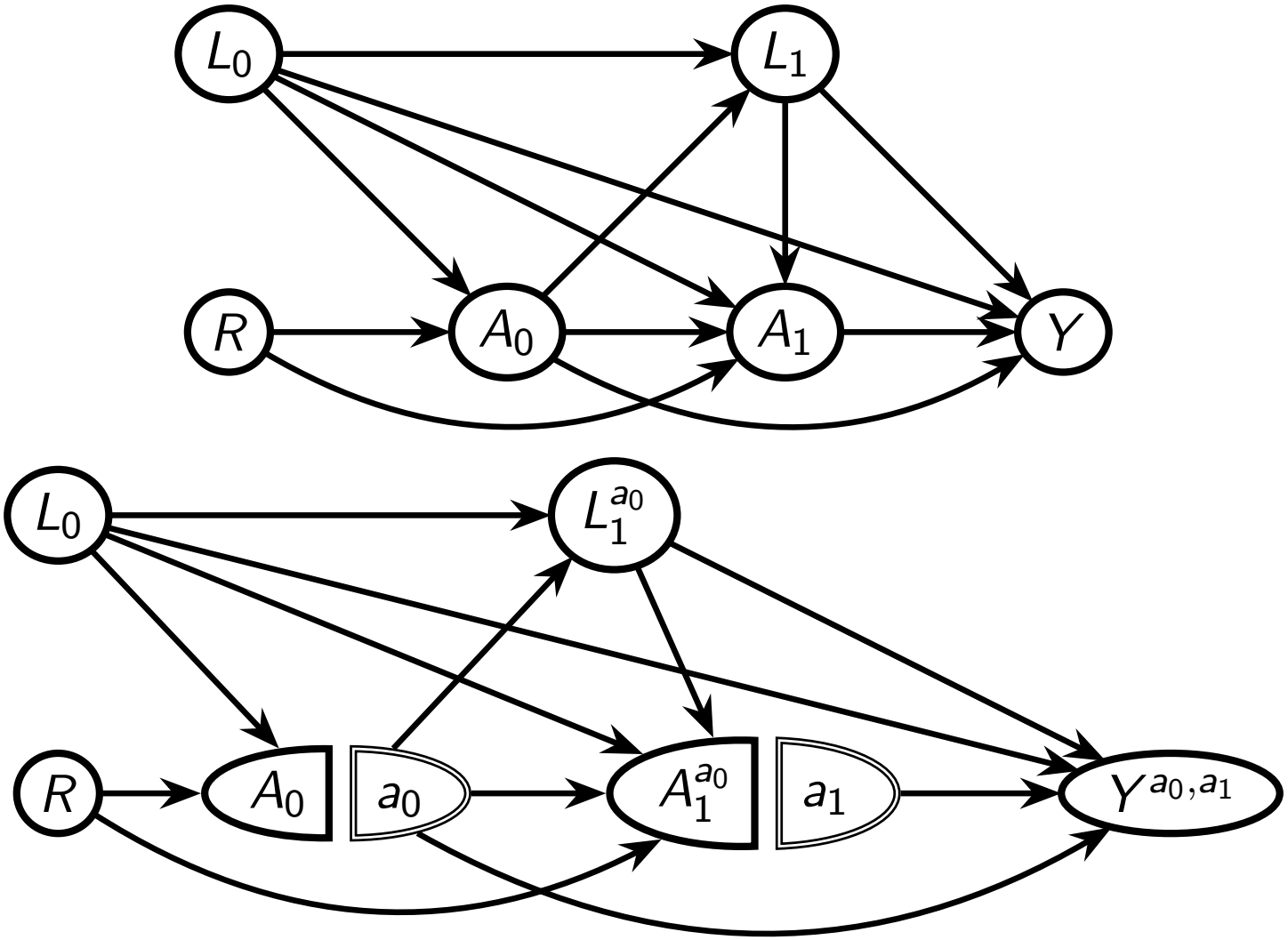
SWIG in an experiment with imperfect adherence

R is the strategy that was assigned, and A_k denotes taking treatment at time $k \in \{0, 1\}$.



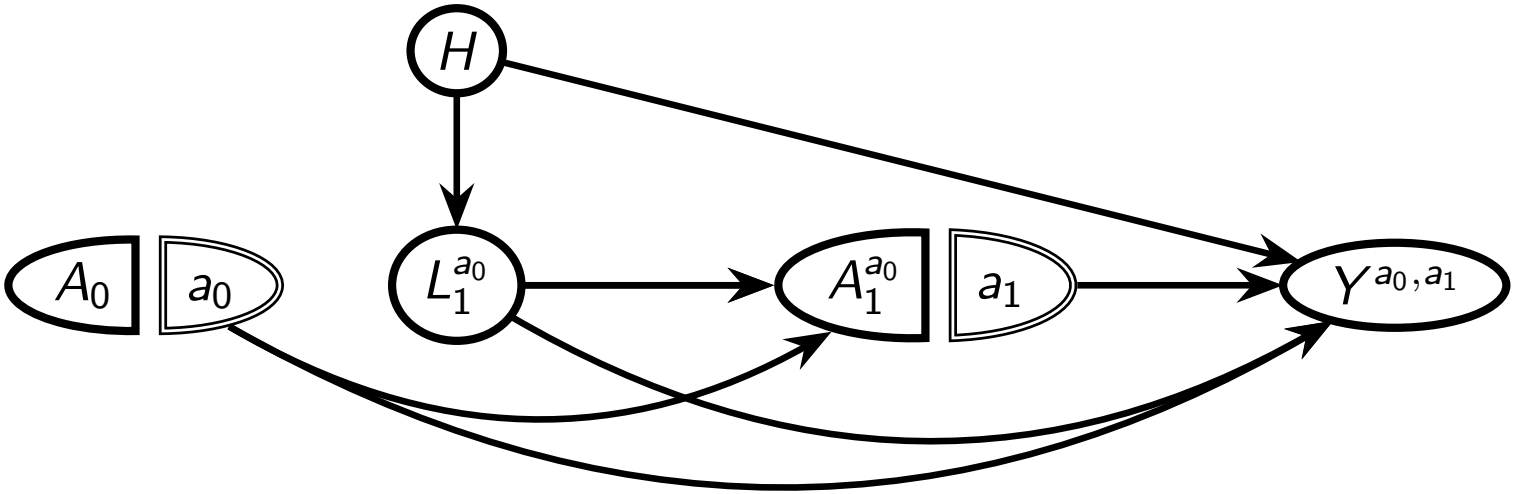
SWIG in an experiment with imperfect adherence

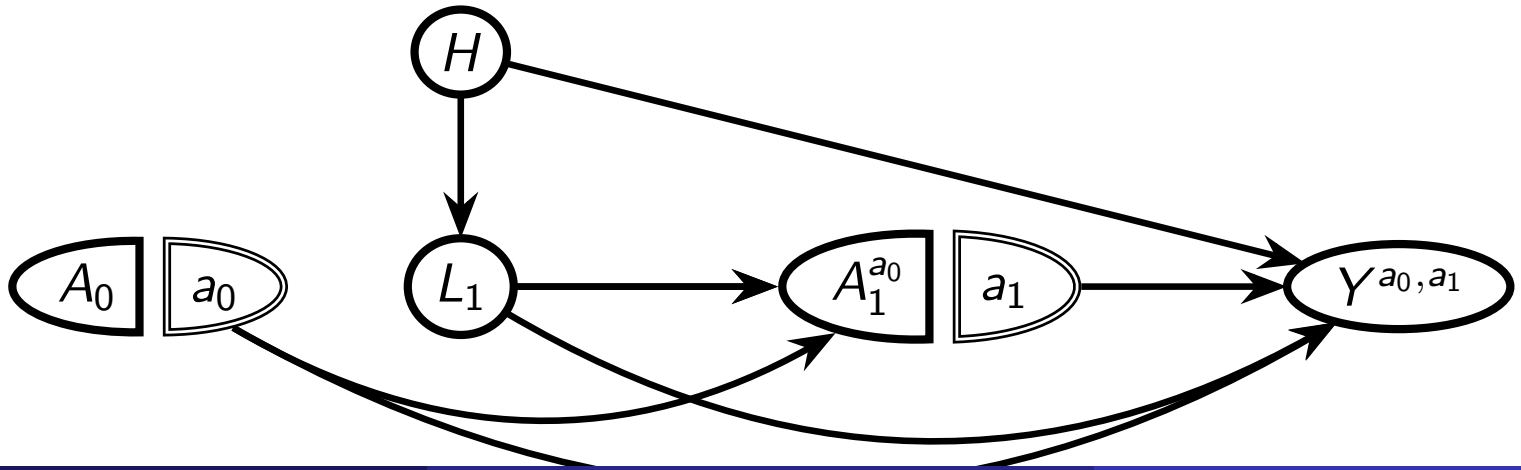
R is the strategy that was assigned, and A_k denotes taking treatment at time $k \in \{0, 1\}$.



SWIG and independencies

These graphs illustrate minimal labelling ($L_1^{a_0} = L_1$). The first graph is not minimally labelled, but encodes the same information as the second graph which is minimally labelled.





SWIG criterion for identification of effects

Consider the observed random variables $\overline{A}_K, \overline{L}_K, Y$.

Definition (g-formula)

The g-formula for the *marginal* of $Y \equiv Y_K$ under treatment assignment $\overline{a} = \overline{a}_K = (a_0, \ldots, a_K)$ is defined as

$$b_{\overline{a}}(y) = \sum_{\overline{l}_K} p(y \mid \overline{l}_K, \overline{a}_K) \prod_{j=0}^{K} p(l_j \mid \overline{l}_{j-1}, \overline{a}_{j-1}),$$

where $\overline{l}_k = (l_0, \ldots, l_k)$, $k \leq K$, are instantiations of **observed** variables $\overline{L}_k = (L_0, \ldots, L_k)$, $k \leq K$.

We define variables with subscript "-1", e.g. L_{-1}, are empty.[26]

[26] Robins, "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect"; Richardson and Robins, "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality".

Sufficient condition for identification

Theorem (Identification of static regimes)

Consider an intervention that sets $\bar{a} = \bar{a}_K = (a_0, \ldots, a_K)$. *Under positivity and consistency,*

$$P(Y^{\bar{a}} = y) = b_{\bar{a}}(y)$$

if for $k \in \{0, \ldots, K\}$

$$Y^{\bar{a}} \perp\!\!\!\perp I(A_k = a_k) \mid L_0, \ldots, L_k, A_0 = a_0, \ldots, A_{k-1} = a_{k-1}.$$

This theorem follows from Robins[27] and Richardson and Robins[28], and is closely related to the backdoor theorem of Pearl[29].
The theorem establishes when we can use the g-formula to identify causal effects.

[27] Robins, "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect".

[28] Richardson and Robins, "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality".

[29] Pearl, "Causal diagrams for empirical research".

Proof in a simple case

Consider the case with two treatments (A_0, A_1) and a binary outcome $Y \in \{0, 1\}$. Suppose that $Y^{a_0,a_1} \perp\!\!\!\perp A_0$ and $Y^{a_0,a_1} \perp\!\!\!\perp A_1 \mid L_1, A_0 = a_0$

Proof.

$$
\begin{aligned}
\mathbb{E}(Y^{a_0,a_1}) &= \mathbb{E}(Y^{a_0,a_1} \mid A_0 = a_0) \text{ exchangeability} \\
&= \sum_{l_1} \mathbb{E}(Y^{a_0,a_1} \mid L_1 = l_1, A_0 = a_0) p(l_1 \mid a_0) \\
&= \sum_{l_1} \mathbb{E}(Y^{a_0,a_1} \mid A_1 = a_1, L_1 = l_1, A_0 = a_0) p(l_1 \mid a_0) \text{ exchangeability} \\
&= \sum_{l_1} \mathbb{E}(Y \mid A_1 = a_1, L_1 = l_1, A_0 = a_0) p(l_1 \mid a_0) \text{ consistency, positivity}
\end{aligned}
$$

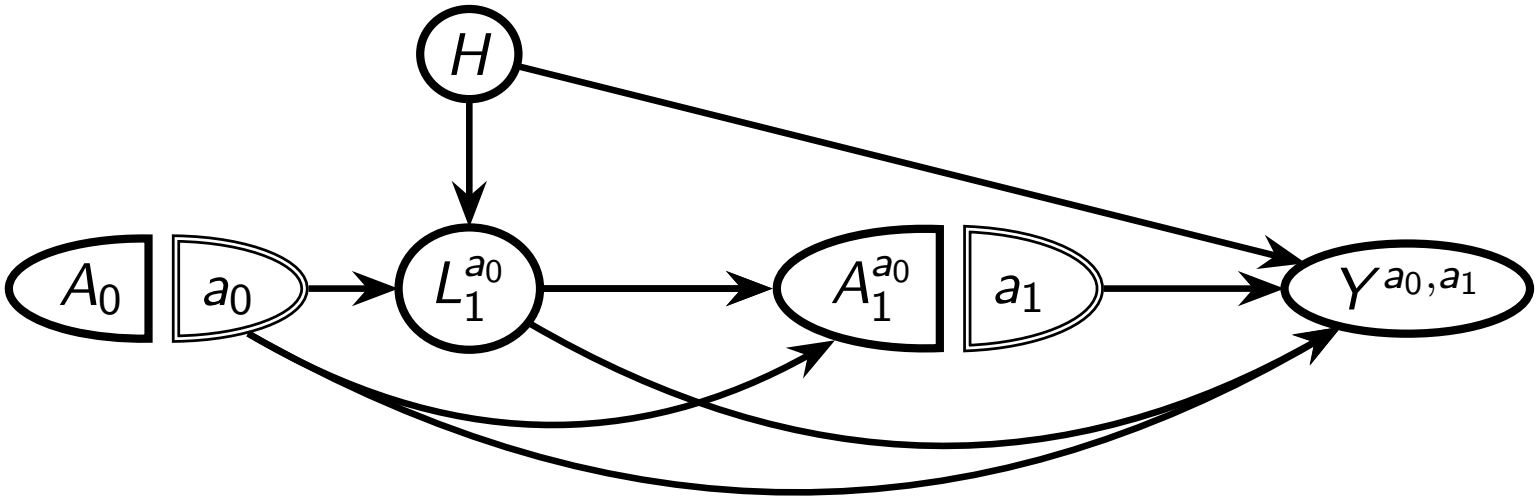
□

Comments to the g-formula

- The independence condition in the identification theorem cannot be read directly off of a SWIG. However, on the next slide we see how the identification condition is implied by an independence in the SWIG.
- Importantly, the g-formula allows identification in the presence of unmeasured variables.

Reading off independencies in SWIGs

Let H be a hidden (unmeasured) variable



We can read off $Y^{a_0,a_1} \perp\!\!\!\perp A_1^{a_0} \mid L_1^{a_0}, A_0$.
However, what we needed for using the g-formula is the independence $Y^{a_0,a_1} \perp\!\!\!\perp A_1 \mid L_1, A_0 = a_0$.
Use consistency: $A_1^{a_0} \mid L_1^{a_0}, A_0 = a_0$ is equal to $A_1 \mid L_1, A_0 = a_0$, i.e.,
$Y^{a_0,a_1} \perp\!\!\!\perp A_1^{a_0} \mid L_1^{a_0}, A_0 \implies Y^{a_0,a_1} \perp\!\!\!\perp A_1 \mid L_1, A_0 = a_0$.

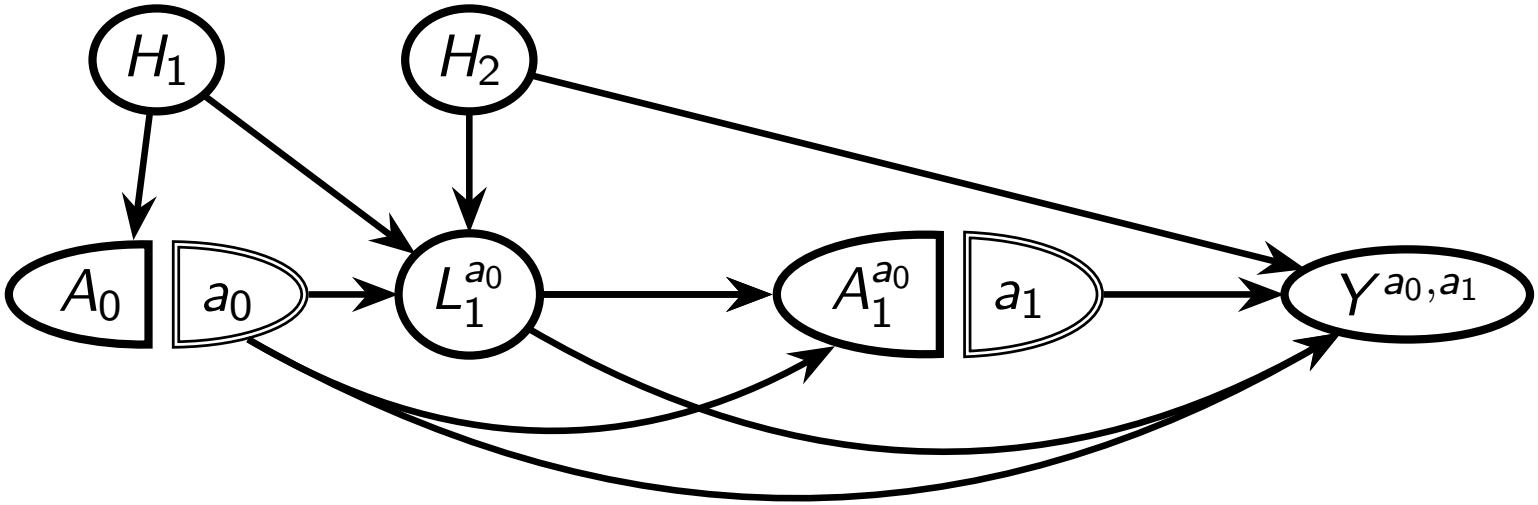
Using the identification theorem

Thus, we can identify the expected counterfactual outcome under the intervention that sets $A_0 = a_0$ and $A_1 = a_1$ in the graph in Slide 189 as

$$\mathbb{E}(Y^{a_0,a_1}) = \sum_{l_1} \mathbb{E}(Y \mid A_1 = a_1, L_1 = l_1, A_0 = a_0) P(L_1 = l_1 \mid A_0 = a_0).$$

Note that we have identified the counterfactual as a function of only the observed variables in the graph, even if there is a hidden variable H in the graph.

Additional SWIG

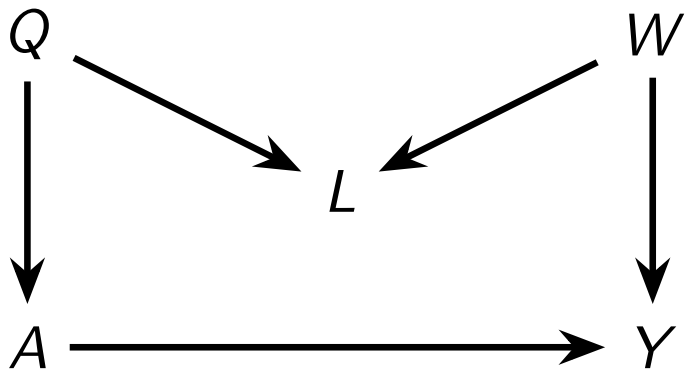


What is the g-formula? Compare to Figure 189. Indeed, the g-formula is just a function of observed data distributions, but here it does not identify the causal estimand because the identification conditions are violated.

Some insights

- We have studied identification from an "all or nothing" perspective.
  - We will later look at sensitivity analyses and bounds.
- The identification assumptions we have studied are non-parametric (PS: I consider this to be a feature, not a bug). We have not considered other assumptions that also can be used to justify identification, for example
  - monotone effects.
  - no effect modification.
- We have **not learned** the graphical structure. On the other hand, we have learned what we can infer from a given graphical structure; heuristically, we encode what we know and believe in the graph, and then we deduce what we can learn from this knowledge and assupmtions.
  - Learning the graphical structure itself from data is a very ambitious task.
  - In principle, the causal structure could be learned by doing a large amount of experiments (I am not discussing this in more detail here).

Berkson bias



- A Drink a glass of red wine a day.
- Y Nausea
- L Aspirin
- Q Family history of cardiovascular disease
- W Frequency of headache

Q: We measure Aspirin. Should we adjust for Aspirin in the analysis? Draw the SWIG...