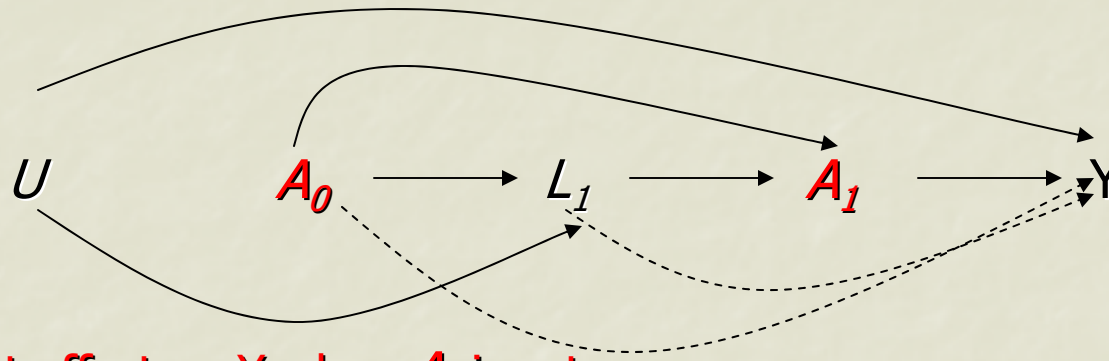


Sequential Randomized Trial of AZT (A_0) and AP (A_1) on survival Y . (L_1 is PCP)



A_0 has no direct effect on Y when A_1 is set.

Arrows from L_1 and A_0 to A_1 represent randomization probability for A_1 depends on A_0 and L_1

No arrows from U directly into A_0 and A_1 due to sequential randomization

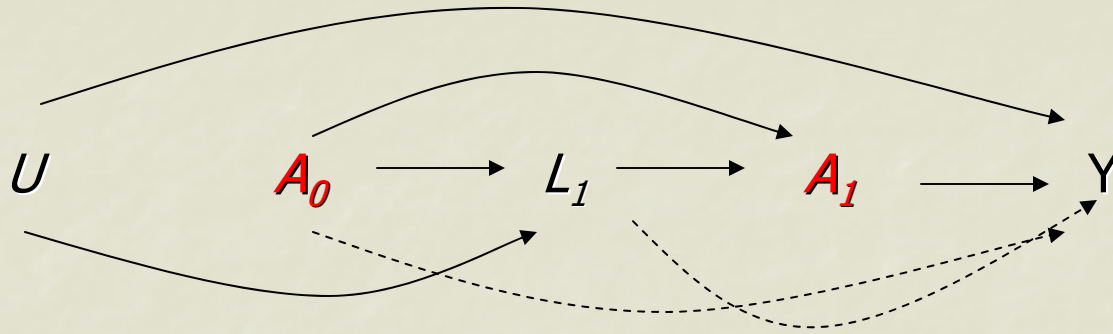
Arrow from A_0 to L_1 represent AZT causes PCP

Arrow from A_1 to Y represent AP causes survival

No direct arrow from A_0 or L_1 to Y represent A_0 does not cause survival except through A_1

Arrows from U to L_1 and Y denote unmeasured immune status determines PCP and survival

Conventional analyses of no direct effects



Test the null hypothesis $E(Y | A_1, A_0 = 1) = E(Y | A_1, A_0 = 0)$

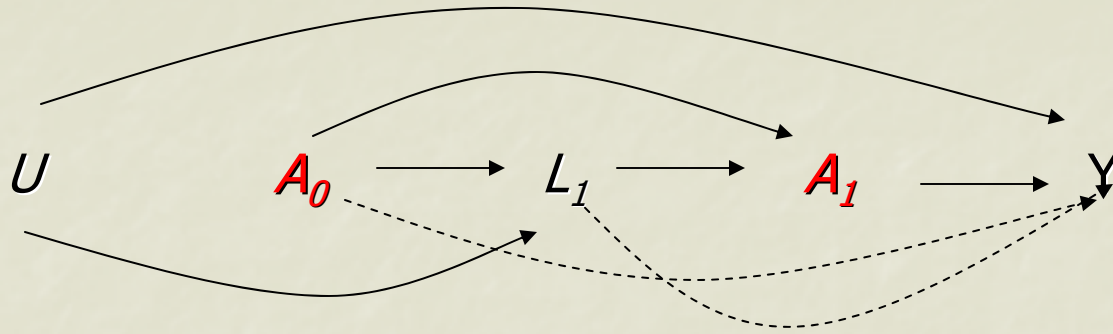
For example, specify the model

$$\text{logit } \Pr(Y = 1 | A_0, A_1) = \beta_0 + \beta_1 A_1 + \beta_2 A_0$$

And test $\beta_2 = 0$

- **However: Even under the causal null hypothesis**
 $\beta_2 \neq 0$ **due to paths $A_0 A_1 L_1 U Y$ and $A_0 L_1 U Y$**

Alternative conventional analysis of no direct effects



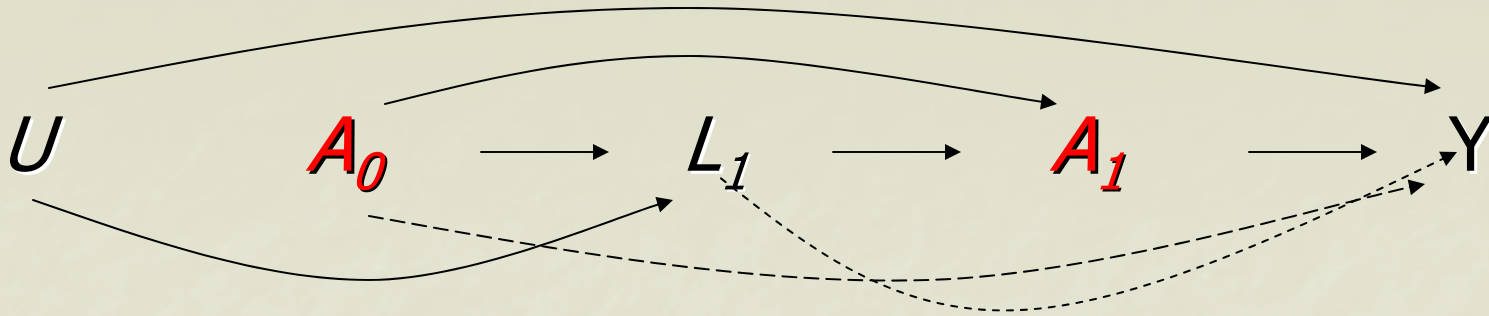
Test the null hypothesis $E(Y | L_1, A_1, A_0 = 1) = E(Y | L_1, A_1, A_0 = 0)$

For example, specify the model

$$\text{logit } \Pr(Y = 1 | A_0, A_1) = \gamma_0 + \gamma_1 A_1 + \gamma_2 A_0 + \gamma_3 L_1$$

And test $\gamma_2 = 0$

- **However: Even under the causal null hypothesis $\gamma_2 \neq 0$ due to path $A_0 L_1 U Y$**



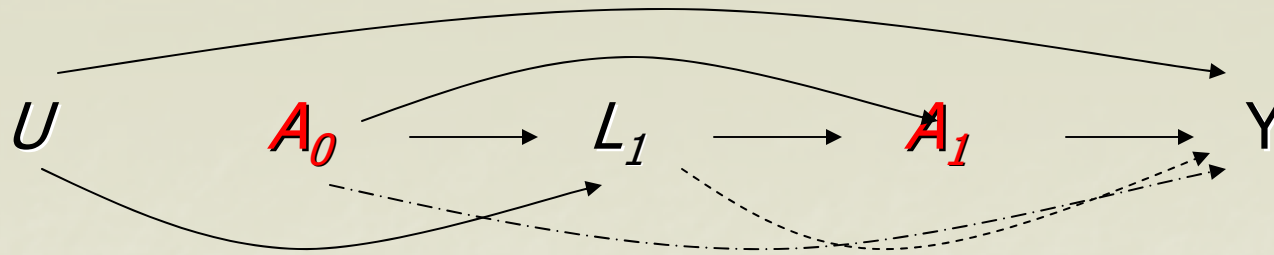
We should be able to test the null hypothesis since we have a sequential randomized trial

Theorem: If U has no arrows directly into treatments, i.e.

$$A_0 \perp\!\!\!\perp U \quad \text{and} \quad A_1 \perp\!\!\!\perp U | L_1, A_0$$

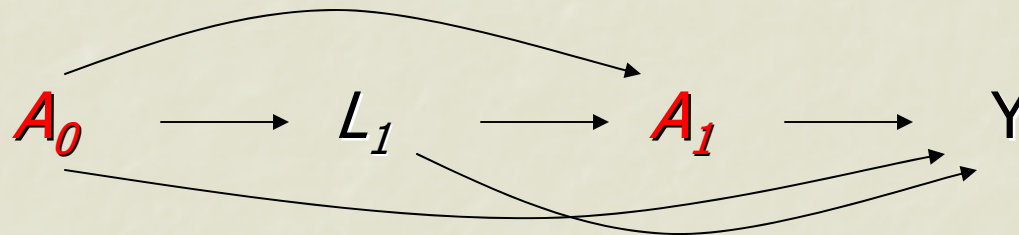
then, U is a non-confounder for the effect of A_0 and A_1 on Y given data on L_1 , that is

$$Y_{a_0, a_1} \perp\!\!\!\perp A_0 \quad \text{and} \quad Y_{a_0, a_1} \perp\!\!\!\perp A_1 | A_0, L_1$$



Thus, $Y_{a_0, a_1} \perp\!\!\!\perp A_0$ and $Y_{a_0, a_1} \perp\!\!\!\perp A_1 | A_0, L_1$

so G-formula based on complete (non-causal) DAG



is causal, that is, $f_{a_0, a_1}(y) = f(Y_{a_0, a_1} = y)$ so a test

that $f_{a_0, a_1}(y) = \sum_{l_1} f(y | a_0, a_1, l_1) f(l_1 | a_0)$ does

not depend on A_0 is a valid test of the null.

$$\Pr[Y=1 | A_1=1, A_0=1] - \Pr[Y=1 | A_1=1, A_0=0] =$$

$$7/12 - 10/16 = -1/24$$

L_1 is a confounder for the effect of A_1 given data on A_0

$$.5 = \Pr[Y=1 | A_1=1, L_1=1, A_0=1]$$

\neq

$$\Pr[Y=1 | A_1=1, L_1=0, A_0=1] = .75$$

$$1 = \Pr[A_1=1 | L_1=1, A_0=1]$$

\neq

$$\Pr[A_1=1 | L_1=0, A_0=1] = .5$$

So must adjust for L_1 to get joint effect of (A_0, A_1)

$$\Pr[Y=1|A_1=1, L_1=1, A_0=1] - \Pr[Y=1|A_1=1, L_1=1, A_0=0] =$$
$$4,000/8,000 - 10,000/16,000 = -1/8$$

L_1 is affected by earlier treatment

$$.5 = \Pr[L_1=1 | A_0=1]$$

\neq

$$\Pr[L_1=0 | A_0=0] = 1$$

So must not adjust for L_1 to get joint effect of (A_0, A_1)

$$\begin{aligned}
E[Y_{a_0=1, a_1=1}] &= f_{a_0=1, a_1=1}(1) = \\
&= \sum_{\ell_1} f(1|\ell_1, A_0=1, A_1=1) f(\ell_1 | A_0=1) \\
&= E[Y|L_1=0, A_0=1, A_1=1] \Pr(L_1=0 | A_0=1) \\
&\quad + E[Y|L_1=1, A_0=1, A_1=1] \Pr(L_1=1 | A_0=1) \\
&= 3/4 \times 1/2 + 1/2 \times 1/2 = 5/8
\end{aligned}$$

$$\begin{aligned}
E[Y_{a_0=0, a_1=1}] &= f_{a_0=0, a_1=1}(1) = \\
&= E[Y|L_1=1, A_0=0, A_1=1] \Pr(L_1=1 | A_0=0) \\
&= 5/8 \times 1 = 5/8
\end{aligned}$$

Done?

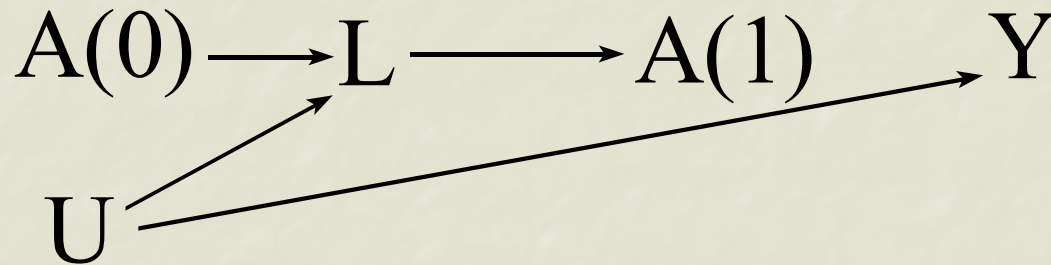
- Unfortunately not
- G-formula, unlike stratification-based methods, is a general method to estimate causal effects but...
- It is non parametric
 - For complex longitudinal data and/or continuous covariates it requires huge amounts of data
 - Computationally intensive
 - No parameter for null hypothesis

Solution

- As usual: Models
- We need models that estimate the same quantities as the g-formula estimates
- Naïve attempt:
 - Plug in standard models to estimate each component of the g-formula
 - Doesn't work

Example:

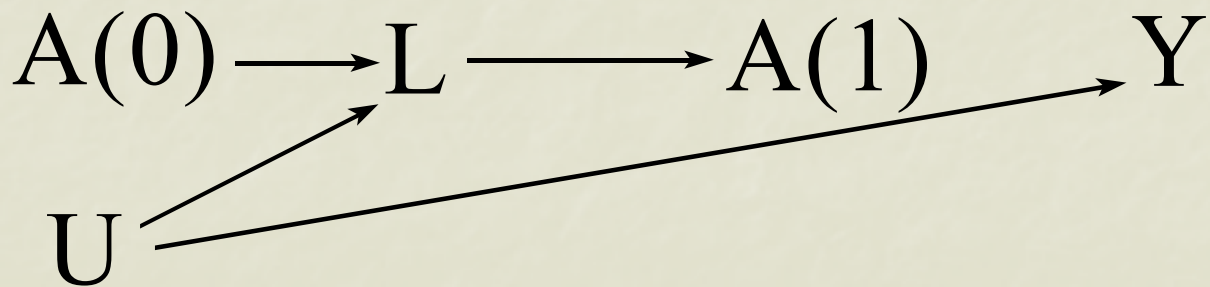
Y continuous, L dichotomous



- $E[Y_{a(0), a(1)}] = \sum_{\ell} E[Y | A(0)=a(0), A(1)=a(1), L=\ell] \Pr[L=\ell | A(0)=a(0)]$
- Linear regression for $E[Y | a(0), a(1), \ell]$
 - $E[Y | a(0), a(1), \ell] = \theta_0 + \theta_1 a(0) + \theta_2 a(1) + \theta_3 \ell$
- Logistic regression for $\Pr[L=\ell | a(0)]$
 - $\text{logit } \Pr[L=1 | a(0)] = \gamma_0 + \gamma_1 a(0)$

$$E[Y_{a(0),a(1)}] \stackrel{?}{=} \hat{\theta}_0 + \hat{\theta}_1 a(0) + \hat{\theta}_2 a(1) + \hat{\theta}_3 \frac{\exp(\hat{\gamma}_0 + \hat{\gamma}_1 a(0))}{1 + \exp(\hat{\gamma}_0 + \hat{\gamma}_1 a(0))}$$

- Calculation does not depend on $a(0), a(1)$ if
 - $\theta_1 = \theta_2 = \theta_3 = 0$ or $\theta_1 = \theta_2 = \gamma_1 = 0$
- $\theta_3 \neq 0$ because $Y \circlearrowleft L \mid A(0), A(1)$
- $\gamma_1 \neq 0$ because $L \circlearrowleft A(0)$



Can't use standard models

- If causal null hypothesis is true, then we know model is misspecified...
before we collect any data!
- We need other kinds of models
- To get there, first we will write the g-formula in a different way...