Causal Inference in Complex Longitudinal Settings

Tutorial on Personalized Medicine, Treatment Regimes, Reinforcement Learning, and Causal Inference

Jon Huang | Singapore Institute for Clinical Sciences | 12 February 2019

Schedule

Time	Topic
09:00 - 09:15 AM	Defining a (longitudinal) causal question
09:15 - 09:30 AM	Implementing the g-formula
09:30 - 10:00 AM	Case Study 1 (smoking cessation)

BREAK (10:00 - 10:30 AM)

Time	Topic
10:30 - 10:55 AM	Case Study 2 (HIV/ART)
10:55 - 11:15 AM	Case Study 3 (aspirin)
11:15 - 11:30 AM	Other considerations, methods, Q&A

Singapore Institute for Clinical Sciences
A*STAR

Logistics

Singapore Institute for Clinical Sciences

Logistics

- Opportunity to follow along with R code

Singapore Institute for Clinical Sciences

Logistics

- Opportunity to follow along with R code

- Will walk through steps to access

Singapore Institute for Clinical Sciences

Logistics

- Opportunity to follow along with R code

- Will walk through steps to access

- Presentation and code are available

Singapore Institute for Clinical Sciences

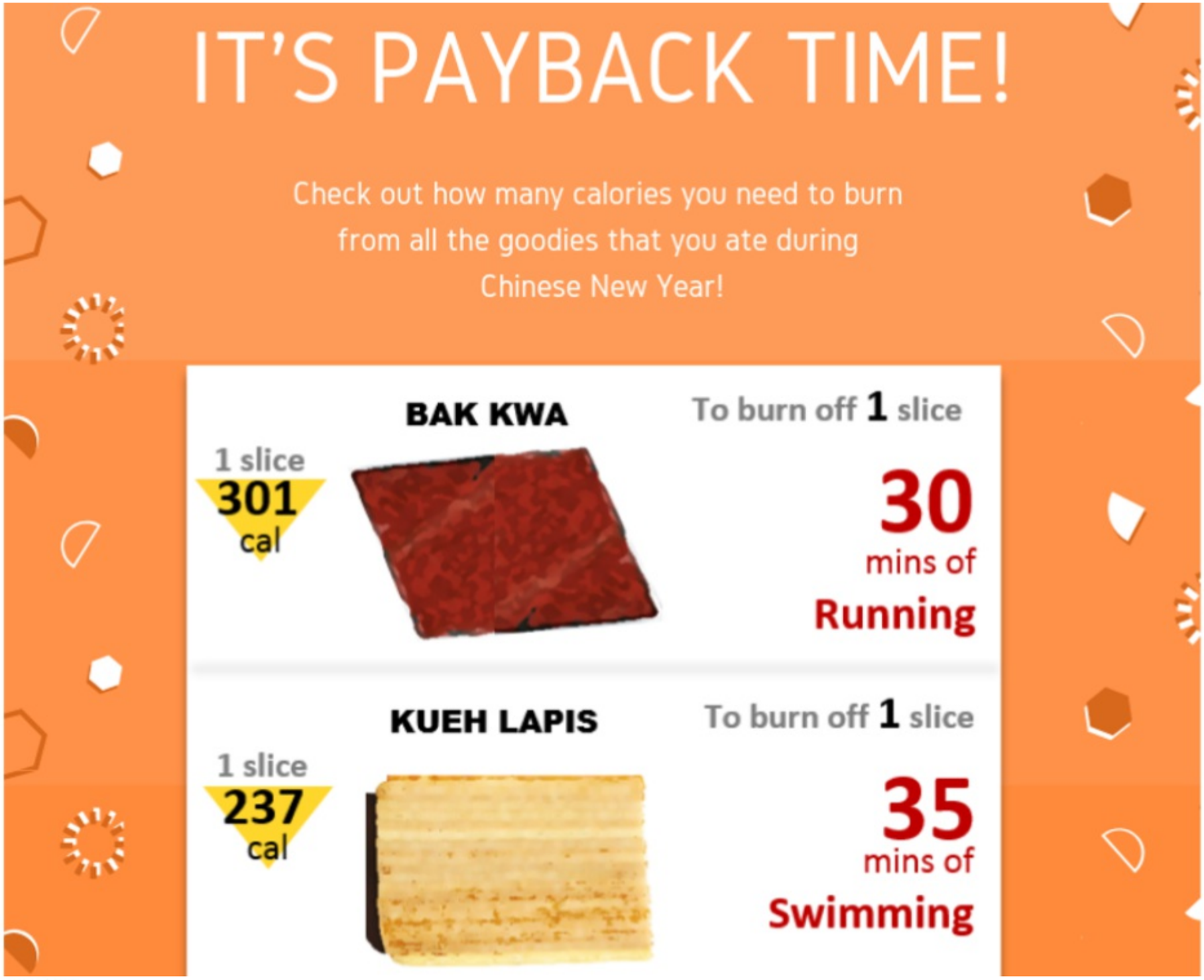
A*STAR

Logistics

- Opportunity to follow along with R code
- Will walk through steps to access
- Presentation and code are available
- Engage and ask questions!

Singapore Institute for Clinical Sciences
A*STAR

What's the question?



Singapore Institute for Clinical Sciences
A*STAR

DEFINING A (LONGITUDINAL) CAUSAL QUESTION

What is a cause? (Review)

What is a cause? (Review)

In counterfactual terms:

- A ***cause*** produces a outcome that is different from what ***would have been*** observed in its absence

Singapore Institute for Clinical Sciences

What is a cause? (Review)

In counterfactual terms:

- A ***cause*** produces a outcome that is different from what ***would have been*** observed in its absence

Causal effects can be defined by difference in "potential outcomes" (PO):

What is a cause? (Review)

In counterfactual terms:

- A ***cause*** produces a outcome that is different from what ***would have been*** observed in its absence

Causal effects can be defined by difference in "potential outcomes" (PO):

- For $X \in (0, 1)$, individual i has two POs:

$$Y_i^{x=1} \text{ or } Y_i^{x=0}$$

Singapore Institute for Clinical Sciences

A*STAR

What is a cause? (Review)

In counterfactual terms:

- A ***cause*** produces a outcome that is different from what ***would have been*** observed in its absence

Causal effects can be defined by difference in "potential outcomes" (PO):

- For $X \in (0, 1)$, individual i has two POs:

$$Y_i^{x=1} \text{ or } Y_i^{x=0}$$

- If all POs could be observed $\rightarrow$ causal effects (contrasts) *e.g.*

Singapore Institute for Clinical Sciences

What is a cause? (Review)

In counterfactual terms:

- A ***cause*** produces a outcome that is different from what ***would have been*** observed in its absence

Causal effects can be defined by difference in "potential outcomes" (PO):

- For $X \in (0, 1)$, individual i has two POs:

$$Y_i^{x=1} \text{ or } Y_i^{x=0}$$

- If all POs could be observed $\rightarrow$ causal effects (contrasts) *e.g.*

$$ACE_{X \rightarrow Y} = E[Y^{x=1}] - E[Y^{x=0}]$$

Singapore Institute for Clinical Sciences

A*STAR

What is a cause? (Review)

In counterfactual terms:

- A ***cause*** produces a outcome that is different from what ***would have been*** observed in its absence

Causal effects can be defined by difference in "potential outcomes" (PO):

- For $X \in (0, 1)$, individual i has two POs:

$$Y_i^{x=1} \text{ or } Y_i^{x=0}$$

- If all POs could be observed $\rightarrow$ causal effects (contrasts) *e.g.*

$$ACE_{X \rightarrow Y} = E[Y^{x=1}] - E[Y^{x=0}]$$

  - Observational data can be considered incomplete subset of POs

Singapore Institute for Clinical Sciences
A*STAR

What is a cause? (Review)

In counterfactual terms:

- A ***cause*** produces a outcome that is different from what ***would have been*** observed in its absence

Causal effects can be defined by difference in "potential outcomes" (PO):

- For $X \in (0, 1)$, individual i has two POs:

$$Y_i^{x=1} \text{ or } Y_i^{x=0}$$

- If all POs could be observed $\rightarrow$ causal effects (contrasts) *e.g.*

$$ACE_{X \rightarrow Y} = E[Y^{x=1}] - E[Y^{x=0}]$$

  - Observational data can be considered incomplete subset of POs
  - Confounding can thus be defined in terms of bias in set of observed POs.

Singapore Institute for Clinical Sciences
A*STAR

Translating observed statistical parameters to causal quantities

For confounders C, binary exposure X, continuous outcome Y:

Singapore Institute for Clinical Sciences

Translating observed statistical parameters to causal quantities

For confounders *C*, binary exposure *X*, continuous outcome *Y*:

Exchangeability:

$$Y^x \perp X = x \mid C$$

Singapore Institute for Clinical Sciences

A*STAR

Translating observed statistical parameters to causal quantities

For confounders C, binary exposure X, continuous outcome Y:

Exchangeability:

$$Y^x \perp X = x \mid C$$

Positivity:

$$0 < Pr(A = a \mid C = c) < 1$$

Singapore Institute for Clinical Sciences

A*STAR

Translating observed statistical parameters to causal quantities

For confounders C, binary exposure X, continuous outcome Y:

Exchangeability:

$$Y^x \perp X = x \mid C$$

Positivity:

$$0 < Pr(A = a \mid C = c) < 1$$

Consistency:

$$\text{if } X = x \text{ then } Y^x = Y$$

Singapore Institute for Clinical Sciences

A*STAR

Translating observed statistical parameters to causal quantities

For confounders C, binary exposure X, continuous outcome Y:

Exchangeability:

$$Y^x \perp X = x \mid C$$

Positivity:

$$0 < Pr(A = a \mid C = c) < 1$$

Consistency:

$$\text{if } X = x \text{ then } Y^x = Y$$

Causal quantity from observed quantities:

$$E(Y \mid X = x, C) = E(Y^x \mid C)$$

Singapore Institute for Clinical Sciences
A*STAR

How to translate a question to an *estimable* causal constrast?

Singapore Institute for Clinical Sciences

A*STAR

How to translate a question to an *estimable* causal constrast?

Can you justify PO distribution is (conditionally) independent from observed exposure level? (correct model specification and *exchangeability*)

Singapore Institute for Clinical Sciences

How to translate a question to an *estimable* causal constrast?

Can you justify PO distribution is (conditionally) independent from observed exposure level? (correct model specification and *exchangeability*)

- Must know enough about the outcome / disease process

Singapore Institute for Clinical Sciences

How to translate a question to an *estimable* causal constrast?

Can you justify PO distribution is (conditionally) independent from observed exposure level? (correct model specification and *exchangeability*)

- Must know enough about the outcome / disease process
- What else influences it?

Singapore Institute for Clinical Sciences

How to translate a question to an *estimable* causal constrast?

Can you justify PO distribution is (conditionally) independent from observed exposure level? (correct model specification and *exchangeability*)

- Must know enough about the outcome / disease process
- What else influences it?
- Are these influences related to exposure?

Singapore Institute for Clinical Sciences

How to translate a question to an *estimable* causal constrast?

Can you justify PO distribution is (conditionally) independent from observed exposure level? (correct model specification and *exchangeability*)

- Must know enough about the outcome / disease process
- What else influences it?
- Are these influences related to exposure?

Do we have enough observational units at each level of exposure (and covariates) to model POs? (*positivity*)

Singapore Institute for Clinical Sciences

How to translate a question to an *estimable* causal constrast?

Can you justify PO distribution is (conditionally) independent from observed exposure level? (correct model specification and *exchangeability*)

- Must know enough about the outcome / disease process
- What else influences it?
- Are these influences related to exposure?

Do we have enough observational units at each level of exposure (and covariates) to model POs? (*positivity*)

Does the effect of exposure vary based on how that level was attained? (*consistency*)

Singapore Institute for Clinical Sciences

How to translate a question to an *estimable* causal constrast?

Can you justify PO distribution is (conditionally) independent from observed exposure level? (correct model specification and *exchangeability*)

- Must know enough about the outcome / disease process
- What else influences it?
- Are these influences related to exposure?

Do we have enough observational units at each level of exposure (and covariates) to model POs? (*positivity*)

Does the effect of exposure vary based on how that level was attained? (*consistency*)

- Critical, varies by target population, and often least considered aspect

Singapore Institute for Clinical Sciences
A*STAR

Example:

General: "What is the effect of smoking on body weight?"

Singapore Institute for Clinical Sciences

A*STAR

Example:

General: "What is the effect of smoking on body weight?"

Better:

- "... among individuals who ever smoked vs. never?"
- "... among individuals who smoked 10 vs. 11 cigarettes per day?"
- "... among individuals who smoked daily for 10 years stopping smoking?"

Example:

General: "What is the effect of smoking on body weight?"

Better:

- "... among individuals who ever smoked vs. never?"
- "... among individuals who smoked 10 vs. 11 cigarettes per day?"
- "... among individuals who smoked daily for 10 years stopping smoking?"

Each question represents a different target population contrast (*consistency*)

Example:

General: "What is the effect of smoking on body weight?"

Better:

- "... among individuals who ever smoked vs. never?"
- "... among individuals who smoked 10 vs. 11 cigarettes per day?"
- "... among individuals who smoked daily for 10 years stopping smoking?"

Each question represents a different target population contrast (*consistency*)

Each may have different set of relevant confounders (*exchangeability*)

Example:

General: "What is the effect of smoking on body weight?"

Better:

- "... among individuals who ever smoked vs. never?"
- "... among individuals who smoked 10 vs. 11 cigarettes per day?"
- "... among individuals who smoked daily for 10 years stopping smoking?"

Each question represents a different target population contrast (*consistency*)

Each may have different set of relevant confounders (*exchangeability*)

Some may not be estimable from available data (*non-positivity*)

What makes the causal question *longitudinal*?

Singapore Institute for Clinical Sciences

What makes the causal question *longitudinal*?

Longitudinal data = repeated observations over time

Singapore Institute for Clinical Sciences

A*STAR

What makes the causal question *longitudinal*?

Longitudinal data = repeated observations over time

1. Exposure changes over time

Singapore Institute for Clinical Sciences

A*STAR

What makes the causal question *longitudinal*?

Longitudinal data = repeated observations over time

1. Exposure changes over time
2. Covariates change over time

Singapore Institute for Clinical Sciences
A*STAR

What makes the causal question *longitudinal*?

Longitudinal data = repeated observations over time

1. Exposure changes over time
2. Covariates change over time
3. Outcome occurs more than once

Singapore Institute for Clinical Sciences

What makes the causal question *longitudinal*?

Longitudinal data = repeated observations over time

1. Exposure changes over time
2. Covariates change over time
3. Outcome occurs more than once
4. Any/all of the above

Singapore Institute for Clinical Sciences
A*STAR

What makes the causal question *longitudinal*?

Longitudinal data = repeated observations over time

1. Exposure changes over time
2. Covariates change over time
3. Outcome occurs more than once
4. Any/all of the above
5. AND you believe such variations are important to your causal question!

Singapore Institute for Clinical Sciences
A*STAR

Longitudinal example:

"What is the effect of smoking on CVD risk, *independent* of its effect on body weight?"

Singapore Institute for Clinical Sciences

A*STAR

Longitudinal example:

"What is the effect of smoking on CVD risk, *independent* of its effect on body weight?"

Does "independent" mean...

- Intervening to fix everyone to the same BMI?
- Intervening to prevent anyone from gaining or losing weight after treatment?
- In another (magical?) world where smoking does not affect BMI?
- Body weight doesn't affect total smoking duration / intensity?

Longitudinal example:

"What is the effect of smoking on CVD risk, *independent* of its effect on body weight?"

Does "independent" mean...

- Intervening to fix everyone to the same BMI?
- Intervening to prevent anyone from gaining or losing weight after treatment?
- In another (magical?) world where smoking does not affect BMI?
- Body weight doesn't affect total smoking duration / intensity?

How these questions are answered $\rightarrow$ target estimand

Singapore Institute for Clinical Sciences

Longitudinal example:

"What is the effect of smoking on CVD risk, *independent* of its effect on body weight?"

Does "independent" mean...

- Intervening to fix everyone to the same BMI?
- Intervening to prevent anyone from gaining or losing weight after treatment?
- In another (magical?) world where smoking does not affect BMI?
- Body weight doesn't affect total smoking duration / intensity?

How these questions are answered $\rightarrow$ target estimand AND determines whether it can be estimated with given data

Singapore Institute for Clinical Sciences

A*STAR

Longitudinal example:

"What is the effect of smoking on CVD risk, *independent* of its effect on body weight?"

Does "independent" mean...

- Intervening to fix everyone to the same BMI?
- Intervening to prevent anyone from gaining or losing weight after treatment?
- In another (magical?) world where smoking does not affect BMI?
- Body weight doesn't affect total smoking duration / intensity?

How these questions are answered $\rightarrow$ target estimand AND determines whether it can be estimated with given data

Draw the DAG!

Singapore Institute for Clinical Sciences
A*STAR

What makes a longitudinal question *complex*?

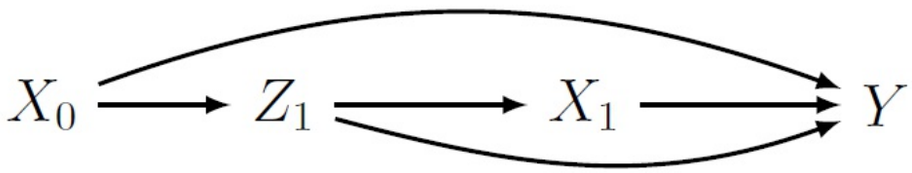
Not complex:

Individual and joint effects can be estimated by conventional methods.

Singapore Institute for Clinical Sciences

A*STAR

What makes a longitudinal question *complex*?



Not complex:

Individual and joint effects can be estimated by conventional methods.

Complex:

Exposure at $t = 1$ is confounded by a consequence of X_0 (Z_1 is a time-varying confounder; exposure-covariate feedback).

Singapore Institute for Clinical Sciences
A*STAR

Complex longitudinal examples:

Complex longitudinal examples:

"What is the effect of taking anti-retroviral therapy (ART) on CD4 counts, taking into account the effect of changes in HIV viral load on subsequent treatment?"

Singapore Institute for Clinical Sciences
A*STAR

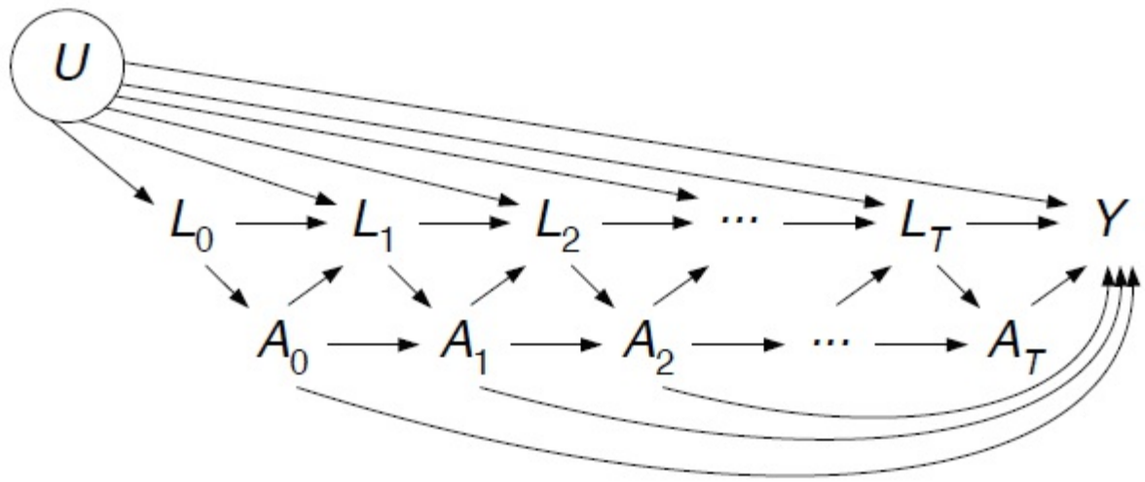
Complex longitudinal examples:

"What is the effect of taking anti-retroviral therapy (ART) on CD4 counts, taking into account the effect of changes in HIV viral load on subsequent treatment?"

"What is the per-protocol effect in a trial of aspirin intake on pregnancy loss, taking into account side effects, non-compliance, and study withdrawal?"

Estimating effects with time-varying treatment and confounding

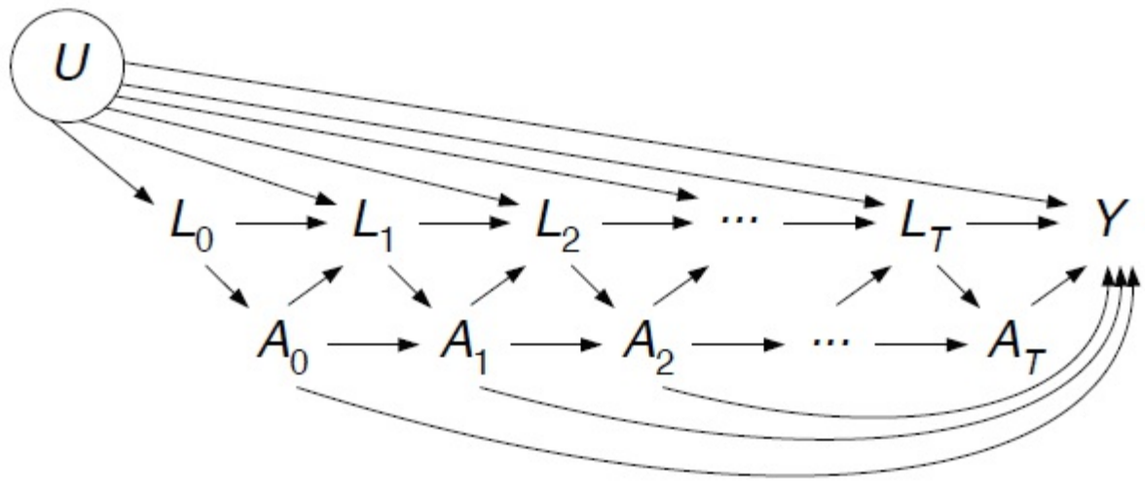
Estimating effects with time-varying treatment and confounding



(Daniel, *et. al. The Stata Journal* 2011.)

Singapore Institute for Clinical Sciences
A*STAR

Estimating effects with time-varying treatment and confounding

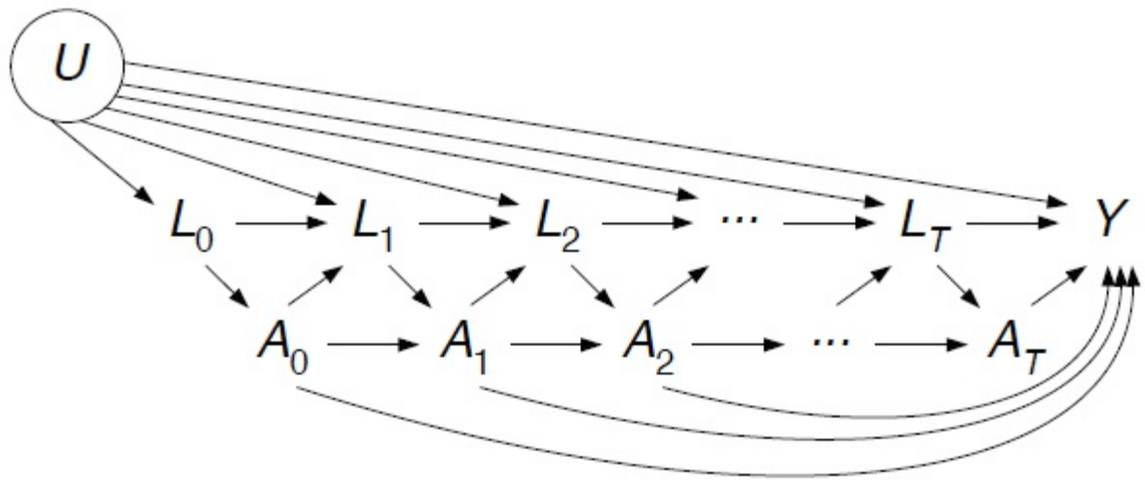


(Daniel, *et. al. The Stata Journal* 2011.)

How would you estimate the causal of effect of A_0?

Singapore Institute for Clinical Sciences
A*STAR

Estimating effects with time-varying treatment and confounding

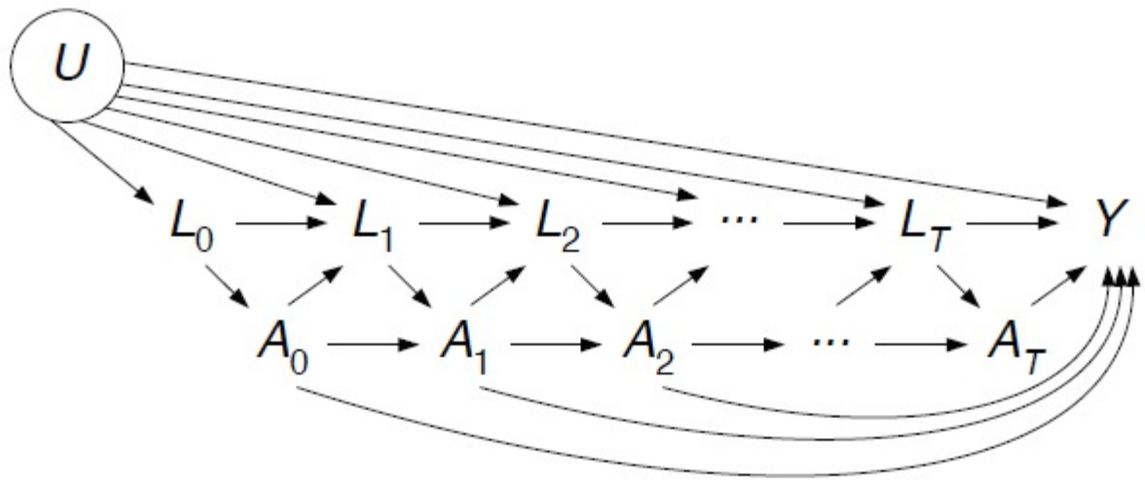


(Daniel, *et. al. The Stata Journal* 2011.)

How would you estimate the causal of effect of A_0? How about A_1?

Singapore Institute for Clinical Sciences
A*STAR

Estimating effects with time-varying treatment and confounding



(Daniel, *et. al. The Stata Journal* 2011.)

How would you estimate the causal of effect of A_0? How about A_1? And overall A?

Singapore Institute for Clinical Sciences
A*STAR

Extension to longitudinal data

For time-varying confounders *L* and binary exposure *A*, continuous outcome *Y*, also consider:

$$\bar{L}_T = l_t, \ldots, l_1, l_0 \text{ and } \bar{A}_T = a_t, \ldots, a_1, a_0 \text{ where } \bar{A}_1\{1,1\} = a_1 = 1, a_0 = 1$$

Singapore Institute for Clinical Sciences

A*STAR

Extension to longitudinal data

For time-varying confounders *L* and binary exposure *A*, continuous outcome *Y*, also consider:

$$\bar{L}_T = l_t, \ldots, l_1, l_0 \text{ and } \bar{A}_T = a_t, \ldots, a_1, a_0 \text{ where } \bar{A}_1\{1,1\} = a_1 = 1, a_0 = 1$$

Sequential exchangeability:

$$Y^{a_1} \perp A_1 = a_1 | L_1, A_0 = a_0$$

Singapore Institute for Clinical Sciences

A*STAR

Extension to longitudinal data

For time-varying confounders *L* and binary exposure *A*, continuous outcome *Y*, also consider:

$$\bar{L}_T = l_t, \ldots, l_1, l_0 \text{ and } \bar{A}_T = a_t, \ldots, a_1, a_0 \text{ where } \bar{A}_1\{1,1\} = a_1 = 1, a_0 = 1$$

Sequential exchangeability:

$$Y^{a_1} \perp A_1 = a_1 | L_1, A_0 = a_0$$

Positivity:

$$0 < Pr(A_1 = a_1 | A_0 = a_0, L_1 = l_1) < 1$$

Singapore Institute for Clinical Sciences

A*STAR

Extension to longitudinal data

For time-varying confounders L and binary exposure A, continuous outcome Y, also consider:

$$\bar{L}_T = l_t, \ldots, l_1, l_0 \text{ and } \bar{A}_T = a_t, \ldots, a_1, a_0 \text{ where } \bar{A}_1\{1,1\} = a_1 = 1, a_0 = 1$$

Sequential exchangeability:

$$Y^{a_1} \perp A_1 = a_1 | L_1, A_0 = a_0$$

Positivity:

$$0 < Pr(A_1 = a_1 | A_0 = a_0, L_1 = l_1) < 1$$

Potential quantities of interest:

$$E(Y | A_1 = a_1, A_0, L_1) = E(Y_1^a | A_0, L_1)$$

Singapore Institute for Clinical Sciences

A*STAR

Extension to longitudinal data

For time-varying confounders L and binary exposure A, continuous outcome Y, also consider:

$$\bar{L}_T = l_t, \ldots, l_1, l_0 \text{ and } \bar{A}_T = a_t, \ldots, a_1, a_0 \text{ where } \bar{A}_1\{1,1\} = a_1 = 1, a_0 = 1$$

Sequential exchangeability:

$$Y^{a_1} \perp A_1 = a_1 | L_1, A_0 = a_0$$

Positivity:

$$0 < Pr(A_1 = a_1 | A_0 = a_0, L_1 = l_1) < 1$$

Potential quantities of interest:

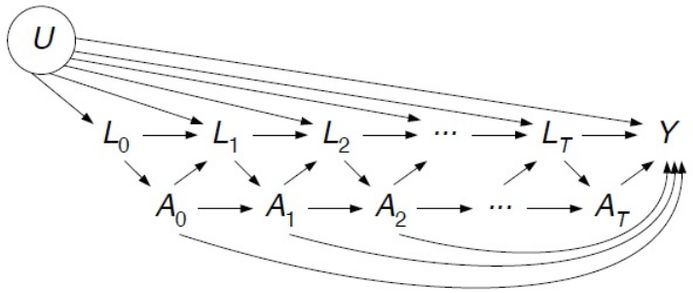
$$E(Y | A_1 = a_1, A_0, L_1) = E(Y_1^a | A_0, L_1)$$

or,

$$E(Y | \bar{A}_1 = 1, \bar{A}_0 = 1, \bar{L}_1) = E(Y^{a_1} | \bar{L}_1)$$

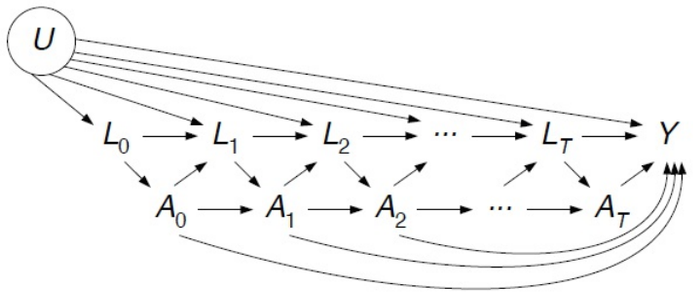
Singapore Institute for Clinical Sciences
A*STAR

BUT!



To estimate joint effect of continous treatment $\bar{A}_1 = \{1,1\}$; confounding by L_1

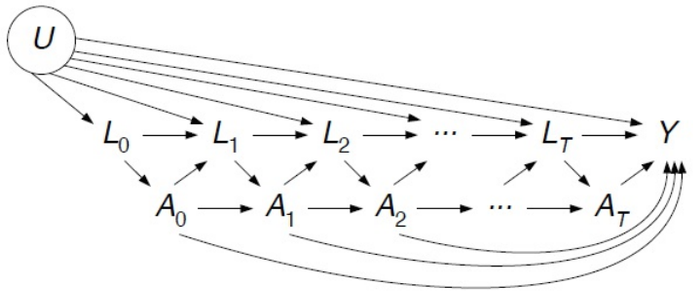
BUT!



To estimate joint effect of continous treatment $\bar{A}_1 = \{1,1\}$; confounding by L_1

e.g. health deteriorates; $\uparrow$ treatment; poorer *observed* outcomes when $A_1 = 1$

BUT!



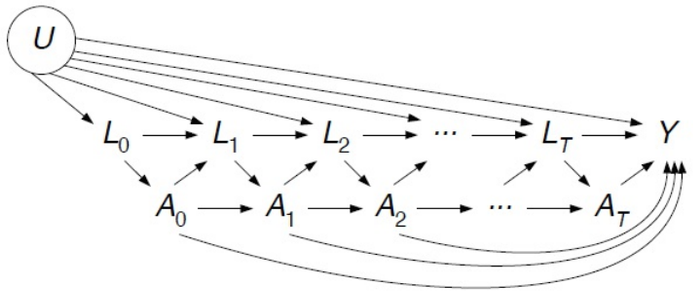
To estimate joint effect of continous treatment $A_1 = \{1,1\}$; confounding by L_1

e.g. health deteriorates; $\uparrow$ treatment; poorer *observed* outcomes when $A_1 = 1$

HOWEVER, standard regression on L_1 induces biasing pathway between A_0 and U and eliminates part of A_0 effect

Singapore Institute for Clinical Sciences

BUT!



To estimate joint effect of continous treatment $A_1 = \{1,1\}$; confounding by L_1

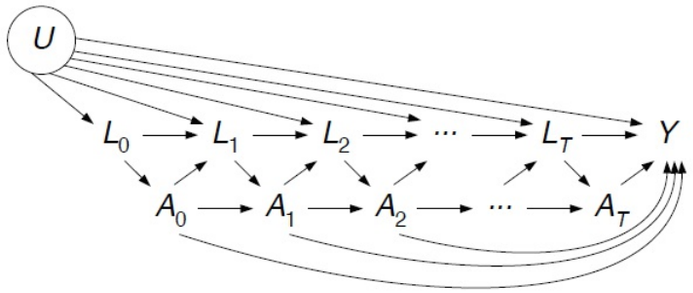
e.g. health deteriorates; $\uparrow$ treatment; poorer *observed* outcomes when $A_1 = 1$

HOWEVER, standard regression on L_1 induces biasing pathway between A_0 and U and eliminates part of A_0 effect

"Adjusting for the future" $\rightarrow$ biased post-treatment strata (collider stratification bias)

Singapore Institute for Clinical Sciences

BUT!



To estimate joint effect of continous treatment $A_1 = \{1,1\}$; confounding by L_1

e.g. health deteriorates; $\uparrow$ treatment; poorer *observed* outcomes when $A_1 = 1$

HOWEVER, standard regression on L_1 induces biasing pathway between A_0 and U and eliminates part of A_0 effect

"Adjusting for the future" $\rightarrow$ biased post-treatment strata (collider stratification bias)

Solution?

Singapore Institute for Clinical Sciences
A*STAR

IMPLEMENTING THE G-FORMULA

Effect estimation under exposure-confounder feedback

True causal effect of A_0 and A_1 = *0*;

$Pr(L_1 | A_0 = 1) = 0.5$, 0.75 untreated; $Pr(A_1 | L_1 = 1) = 0.8$, 0.4 otherwise

$E[Y^{a_0=1}] =$ 76*(48/160)* + 76(32/160) + 44*(16/160)* + *44* (64/160) = ***60***

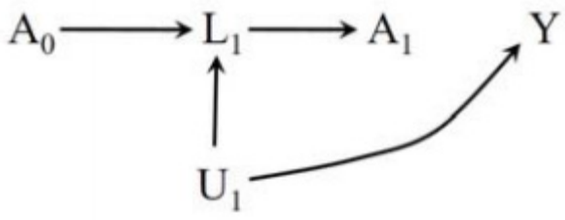
$E[Y^{a_0=0}] =$ 84*(24/160)* + *84*(16/160) + 52*(24/160)* + *52* (96/160) = ***60***

$E[Y^{a_0=1}] - E[Y^{a_0=0}] =$ ***0***

Also,

$E[Y^{a_1=1} | A_0, L_1] - E[Y^{a_1=0} | A_0, L_1] =$ ***0***

HOWEVER, joint effect...



N	A_0	L_1	A_1	Mean Y
2400	0	0	0	84
1600	0	0	1	84
2400	0	1	0	52
9600	0	1	1	52
4800	1	0	0	76
3200	1	0	1	76
1600	1	1	0	44
6400	1	1	1	44

Hernán & Robins (2019)

Singapore Institute for Clinical Sciences
A*STAR

Effect under continuous treatment

Unadjusted (for L_1):

Singapore Institute for Clinical Sciences

Effect under continuous treatment

Unadjusted (for L_1):

$$E[Y|A_0 = 1, A_1 = 1] = 76(32/96) + 44(64/96) = 54.7 \quad E[Y|A_0 = 0, A_1 = 0] = 84(24/48) + 52(24/48) = 68$$

$$E[Y|A_0 = 1, A_1 = 1] - E[Y|A_0 = 0, A_1 = 0] = \textbf{\textit{-13.3}}$$

Singapore Institute for Clinical Sciences

A*STAR

Effect under continuous treatment

Unadjusted (for L_1):

$$E[Y|A_0 = 1, A_1 = 1] = 76(32/96) + 44(64/96) = 54.7 \quad E[Y|A_0 = 0, A_1 = 0] = 84(24/48) + 52(24/48) = 68$$

$$E[Y|A_0 = 1, A_1 = 1] - E[Y|A_0 = 0, A_1 = 0] = \textbf{\textit{-13.3}}$$

Adjust for L_1 by stratification:

Singapore Institute for Clinical Sciences

Effect under continuous treatment

Unadjusted (for L_1):

$$E[Y|A_0 = 1, A_1 = 1] = 76(32/96) + 44(64/96) = 54.7 \quad E[Y|A_0 = 0, A_1 = 0] = 84(24/48) + 52(24/48) = 68$$

$$E[Y|A_0 = 1, A_1 = 1] - E[Y|A_0 = 0, A_1 = 0] = \textbf{\textit{-13.3}}$$

Adjust for L_1 by stratification:

Within $L_1 = 0$:

$$E[Y|A_0 = 1, A_1 = 1, L_1 = 0] - E[Y|A_0 = 0, A_1 = 0, L_1 = 0] = 76 - 84 = \textbf{\textit{-8}}$$

(Subjects where $A_0 = 0$, $L_1 = 0$ healthier than a random draw of subjects)

Within $L_1 = 1$:

$$E[Y|A_0 = 1, A_1 = 1, L_1 = 1] - E[Y|A_0 = 0, A_1 = 0, L_1 = 1] = 44 - 52 = \textbf{\textit{-8}}$$

(Subjects where $A_0 = 1$, $L_1 = 1$ sicker than a random draw of subjects)

Stratification is the issue!

Singapore Institute for Clinical Sciences
A*STAR

Standardization to estimate counterfactual means

For point exposure and time-fixed confounders:

$$E(Y) = \sum_{A} \sum_{L} E(Y \mid A, L) P(A \mid L) P(L)$$

Singapore Institute for Clinical Sciences

A*STAR

Standardization to estimate counterfactual means

For point exposure and time-fixed confounders:

$$E(Y) = \sum_A \sum_L E(Y \mid A, L)P(A \mid L)P(L)$$

If our model is correct, and causal assumptions hold:

$$E[Y^a] = \sum_l E[Y \mid A = a, L = l] * f(l), \text{ where } f(l) = Pr(L = l)$$

i.e. weighted conditional mean of Y standardized to observed covariate distribution

Singapore Institute for Clinical Sciences

A*STAR

Standardization to estimate counterfactual means

For point exposure and time-fixed confounders:

$$E(Y) = \sum_A \sum_L E(Y \mid A, L)P(A \mid L)P(L)$$

If our model is correct, and causal assumptions hold:

$$E[Y^a] = \sum_l E[Y \mid A = a, L = l] * f(l), \text{ where } f(l) = Pr(L = l)$$

i.e. weighted conditional mean of *Y* standardized to observed covariate distribution

Extention to time-varying exposures and confounders:

$$E[Y^{a_0, a_1}] = \sum_{l_1} E[Y \mid A_0 = a_0, A_1 = a_1, L_1 = l_1] * f(l_1 \mid a_0), \text{ where } f(l_1 \mid a_0) = Pr(L_1 = l \mid A_0 = a)$$

i.e., weighted conditional mean of *Y* standardized to exposure and covariate history *as they would have been observed under sequential randomization*

Singapore Institute for Clinical Sciences
A*STAR

Applied to the example:

N	A_0	L_1	A_1	Mean Y
2400	0	0	0	84
1600	0	0	1	84
2400	0	1	0	52
9600	0	1	1	52
4800	1	0	0	76
3200	1	0	1	76
1600	1	1	0	44
6400	1	1	1	44

Singapore Institute for Clinical Sciences

A*STAR

Applied to the example:

N	A_0	L_1	A_1	Mean Y
2400	0	0	0	84
1600	0	0	1	84
2400	0	1	0	52
9600	0	1	1	52
4800	1	0	0	76
3200	1	0	1	76
1600	1	1	0	44
6400	1	1	1	44

$E[Y^{a_0=1,a_1=1}] =$

$E[Y|A_0 = 1, A_1 = 1, L_1 = 0] * Pr[L_1 = 0|A_0 = 1] +$
$E[Y|A_0 = 1, A_1 = 1, L_1 = 1] * Pr[L_1 = 1|A_0 = 1]$

= 76 *(80/160)* + *44* (80/160) = ***60***

$E[Y^{a_0=0,a_1=0}] =$

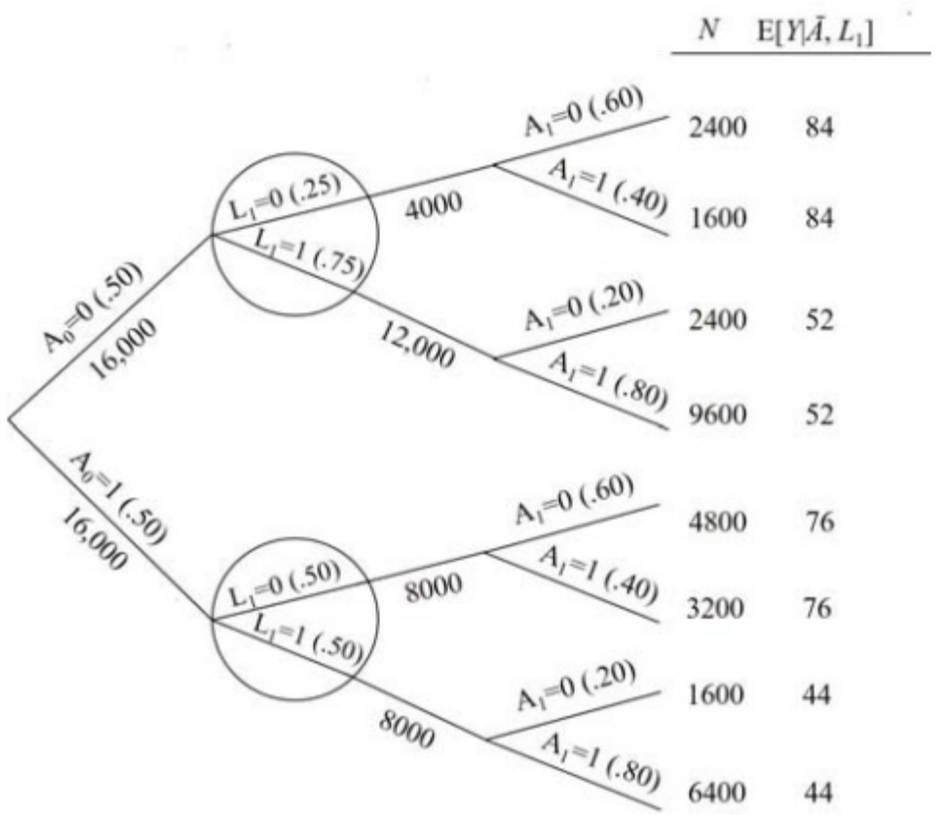
$E[Y|A_0 = 0, A_1 = 0, L_1 = 0] * Pr[L_1 = 0|A_0 = 0] +$
$E[Y|A_0 = 0, A_1 = 0, L_1 = 1] * Pr[L_1 = 1|A_0 = 0]$

= 84 *(40/160)* + *52* (120/160) = ***60***

$E[Y^{a_0=1,a_1=1}] - E[Y^{a_0=0,a_1=0}]$ = ***0***

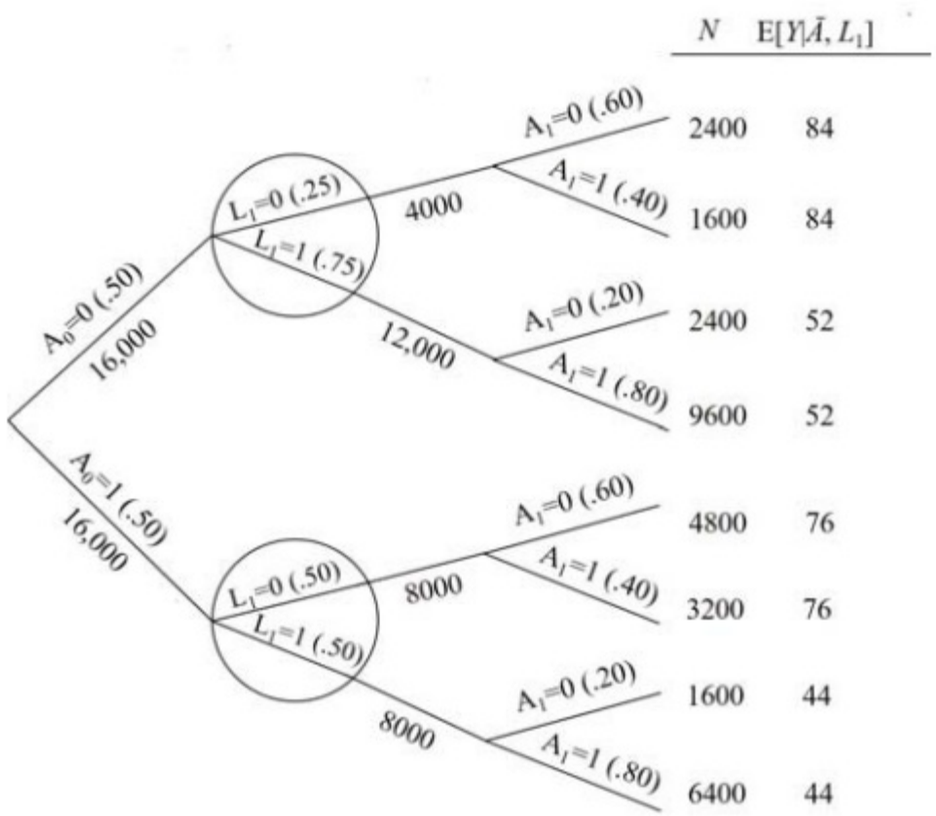
Singapore Institute for Clinical Sciences
A*STAR

Considered as a simulation:

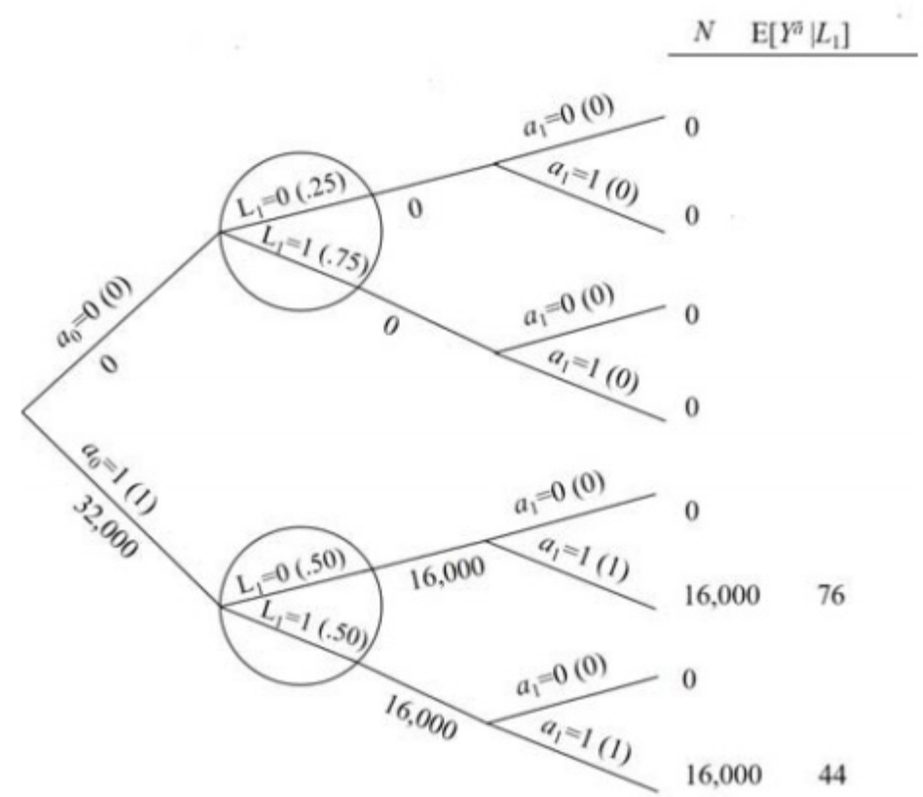


Observed Data

Considered as a simulation:



Observed Data



Simulated

Singapore Institute for Clinical Sciences
A*STAR

Parametric g-formula

$$\sum_{\bar{l}} E[Y \mid \bar{A} = \bar{a}, \bar{L} = \bar{l}] * \prod_{k=0}^{K} f(l_k \mid \bar{a}_{k-1}, \bar{l}_{k-1})$$

where, *e.g.* conditional means estimated by linear, and distribution of discrete covariates by logistic, regressions

"plug-in g-formula" or "parametric g-formula"

can be generalized further to densities

Singapore Institute for Clinical Sciences

A*STAR

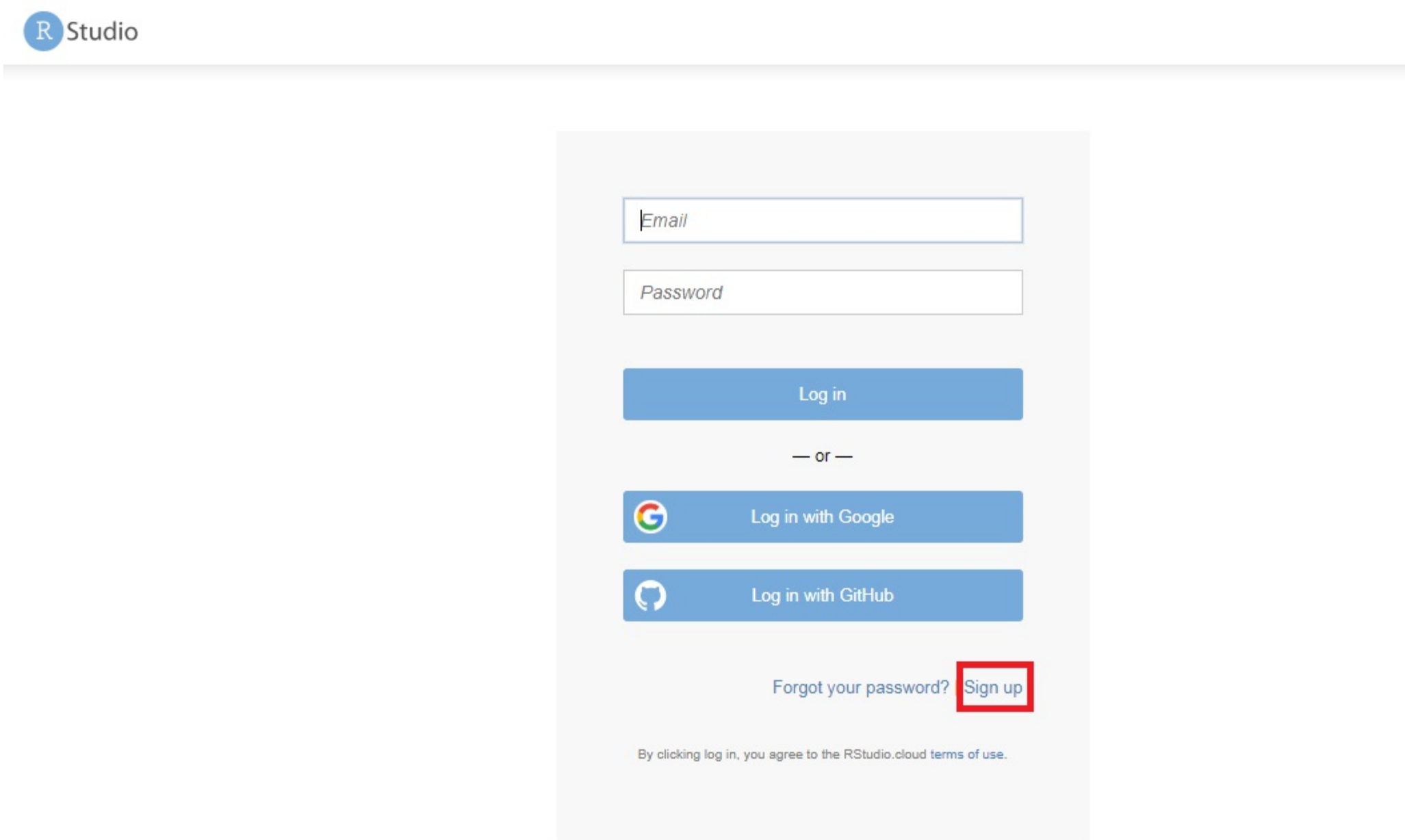
Estimation Steps In Practice

- Draw a DAG representing causal structure
- Model temporal / structural relationships (Q-model) implied by DAG
- Sample baseline values simulate POs using model
- Repeat setting exposures to levels reflecting desired contrast
- Bootstrap to obtain SE

INTERLUDE:

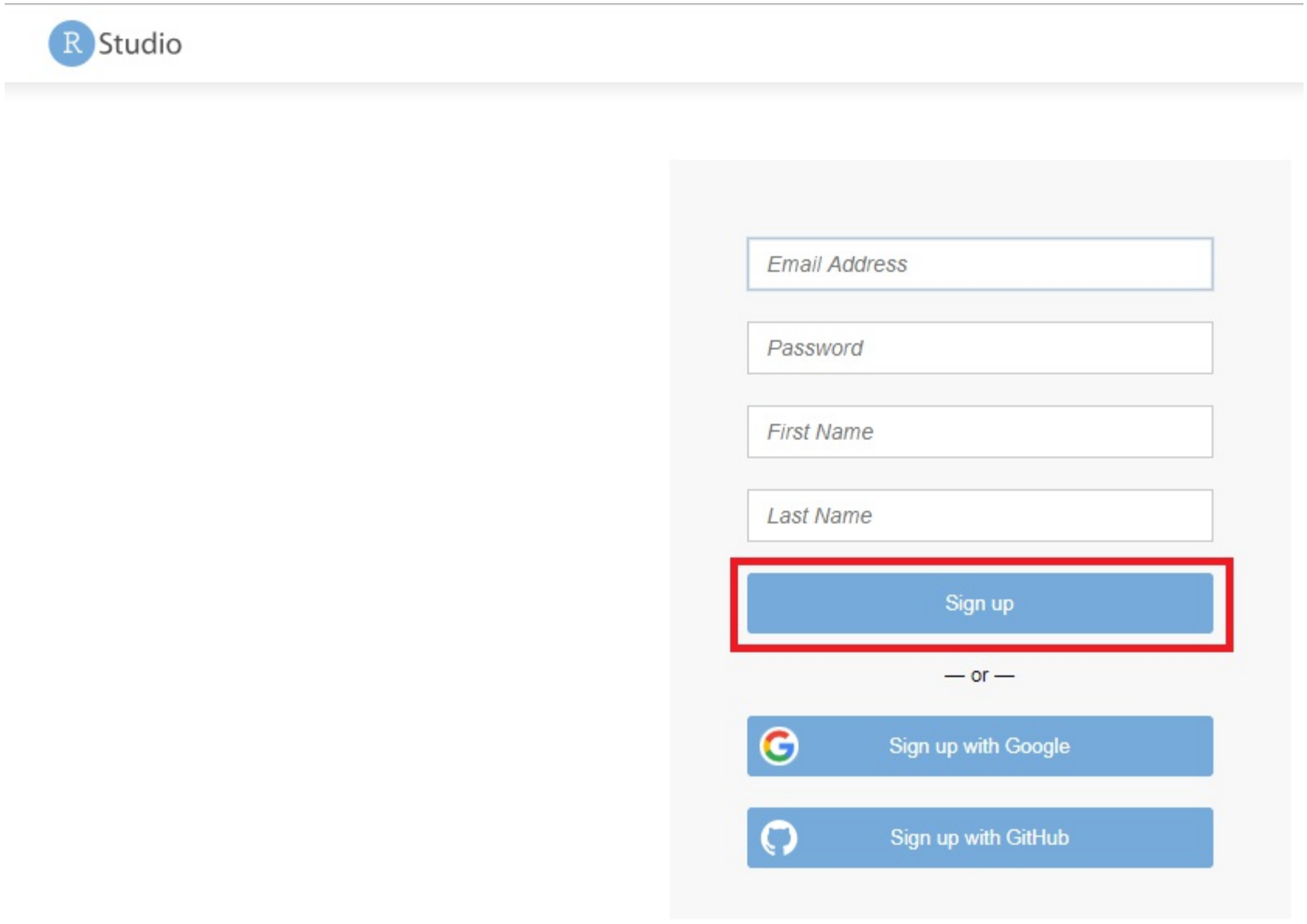
RStudio Cloud Setup

1. bit.ly/CI_tutorial_cloud



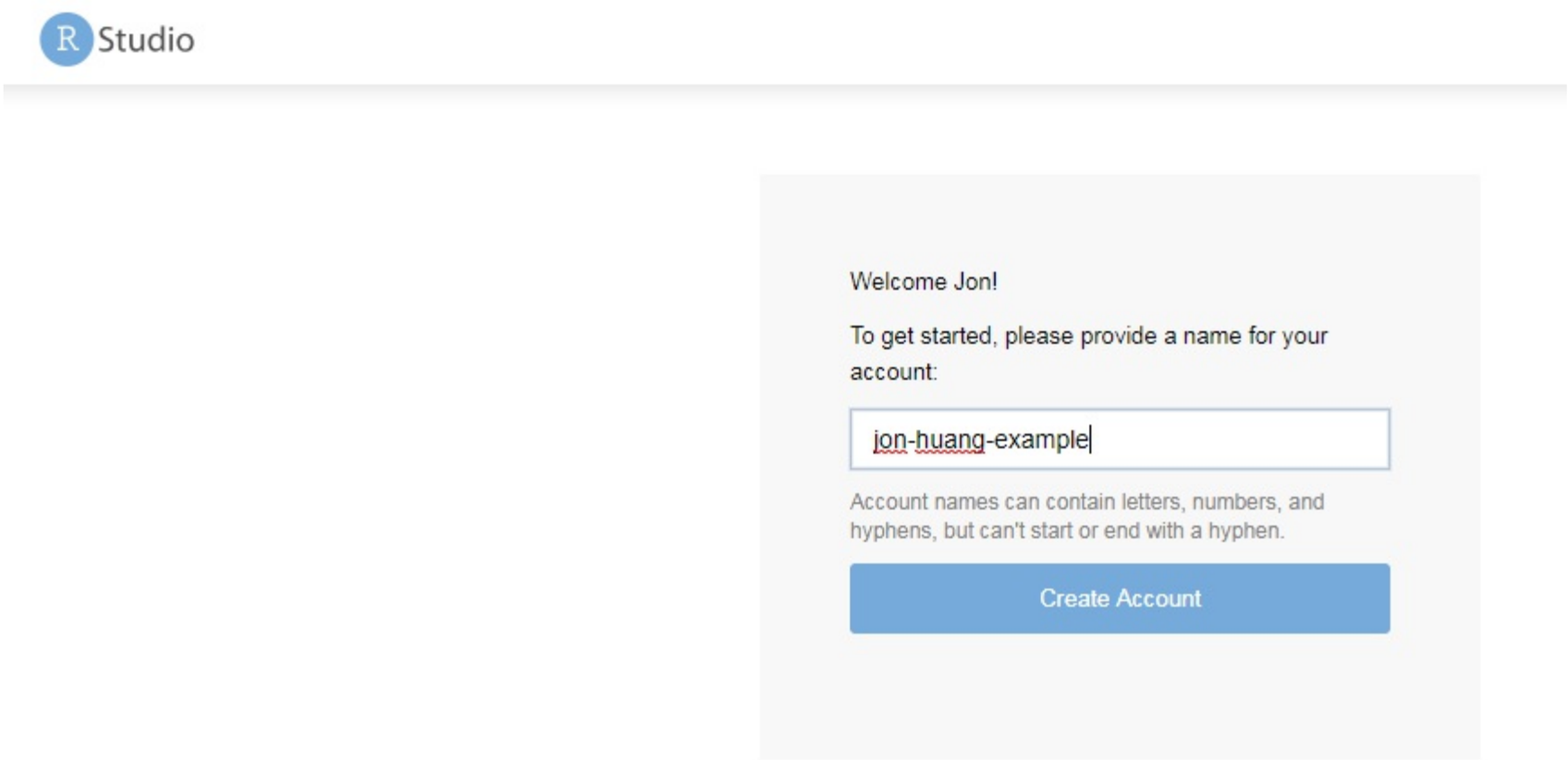
Singapore Institute for Clinical Sciences
A*STAR

1. Create Account



Singapore Institute for Clinical Sciences
A*STAR

1. Create Account



Singapore Institute for Clinical Sciences
A*STAR

2. Navigate to project

NUS-IMS CI Tutorial Projects Members Info Jon Huang

Welcome to NUS-IMS CI Tutorial

12 February 2019

"Introduction to Causal Inference in Complex Longitudinal Settings" Session
Jonathan Y Huang, MPH, PhD

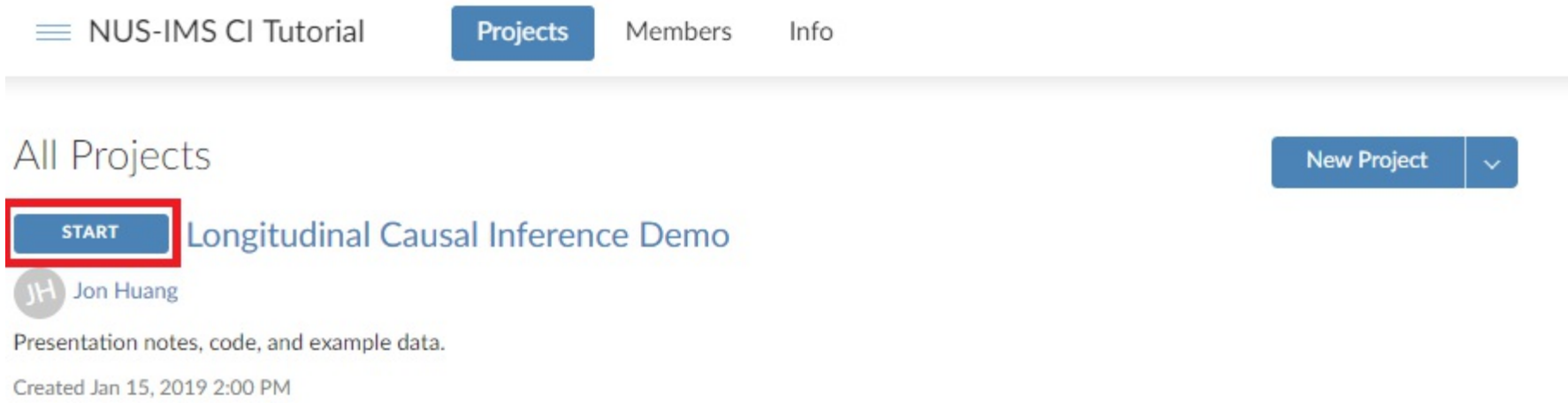
NUS-IMS "Tutorial on Personalized Medicine, Treatment Regimes, Reinforcement Learning, and Causal Inference"

Lecture notes and practice code and data will be accessible here.

Navigate to "bit.ly/CI_tutorial_cloud" to access!

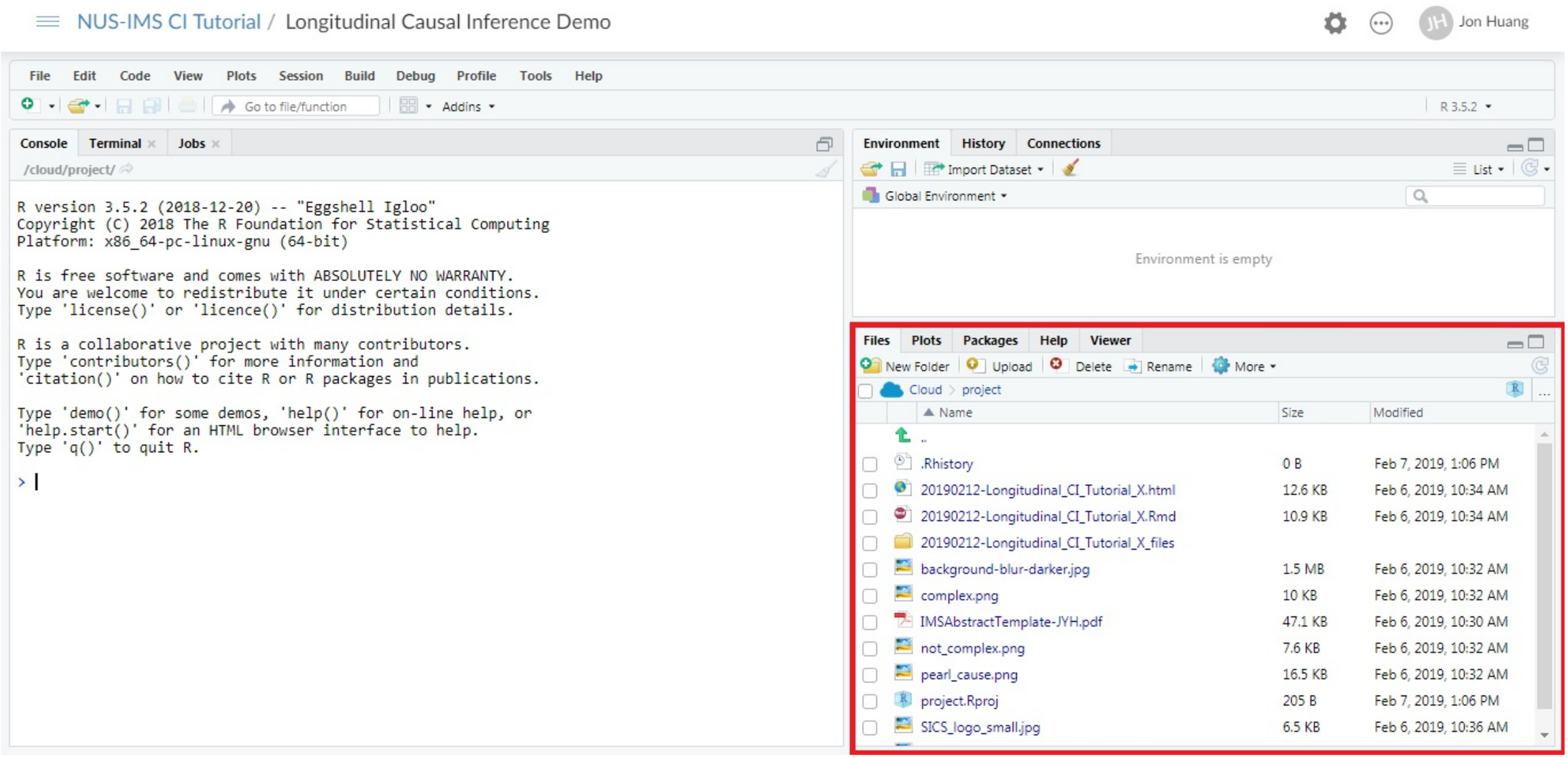
Singapore Institute for Clinical Sciences
A*STAR

2. Navigate to project



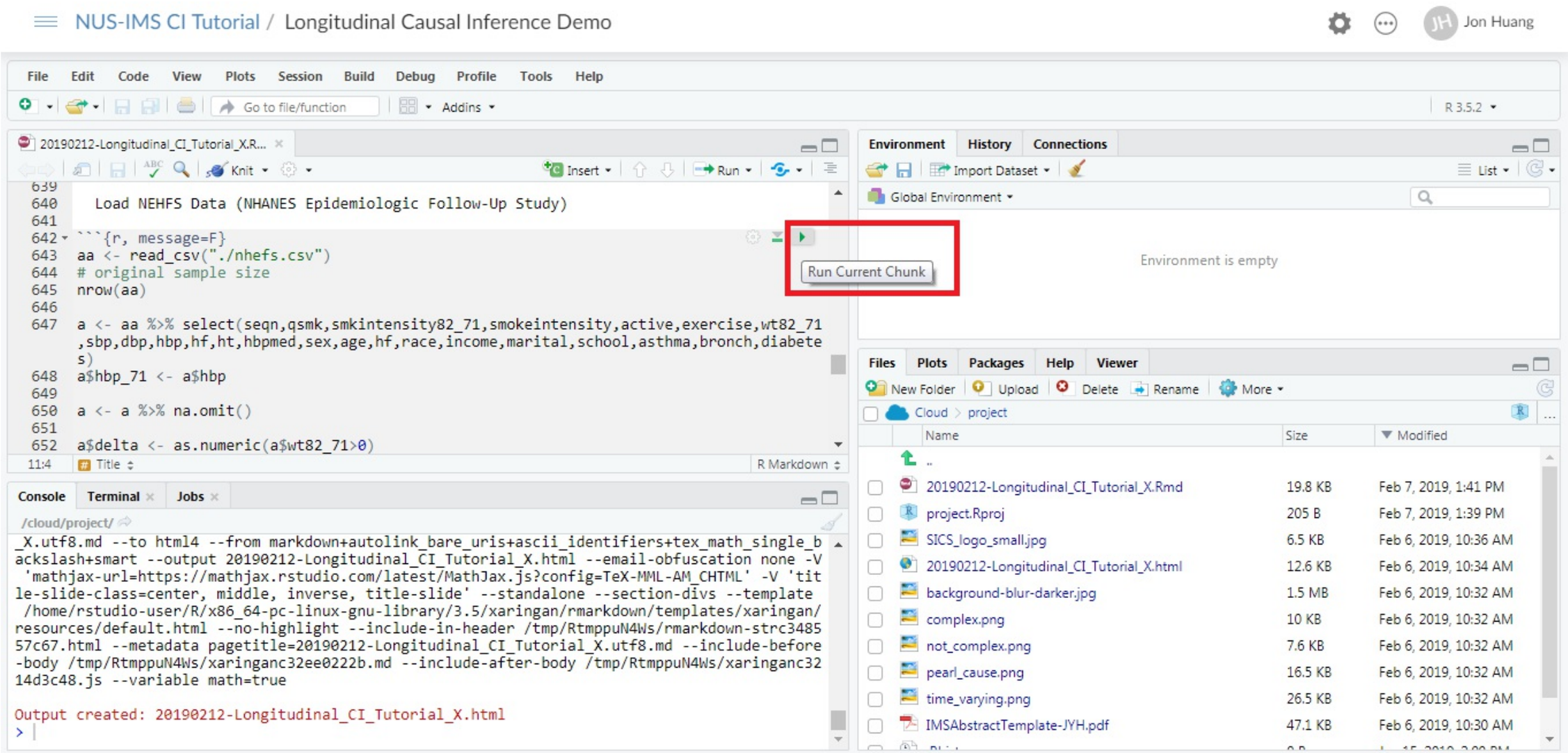
Singapore Institute for Clinical Sciences
A*STAR

3. Load files



Singapore Institute for Clinical Sciences
A*STAR

Two ways to follow along:



1. Easy Way: Scroll to line 700. Click "Play" buttons.

Singapore Institute for Clinical Sciences
A*STAR

Two ways to follow along:

NUS-IMS CI Tutorial / Longitudinal Causal Inference Demo

Jon Huang

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | R 3.5.2

20190212-Longitudinal_CI_Tutorial_X.R...

Knit | Insert | Run

```
639
640   Load NEHFS Data (NHANES Epidemiologic Follow-Up Study)
641
642 ```{r, message=F}
643 aa <- read_csv("./nhefs.csv")
644 # original sample size
645 nrow(aa)
646
647 a <- aa %>% select(seqn,qsmk,smkintensity82_71,smokeintensity,active,exercise,wt82_71
    ,sbp,dbp,hbp,hf,ht,hbpmed,sex,age,hf,race,income,marital,school,asthma,bronch,diabete
    s)
648 a$hbp_71 <- a$hbp
649
650 a <- a %>% na.omit()
651
652 a$delta <- as.numeric(a$wt82_71>0)
```

Run Current Chunk

11:4 | Title | R Markdown

Console | Terminal | Jobs

/cloud/project/

```
_X.utf8.md --to html4 --from markdown+autolink_bare_uris+ascii_identifiers+tex_math_single_b
ackslash+smart --output 20190212-Longitudinal_CI_Tutorial_X.html --email-obfuscation none -V
 'mathjax-url=https://mathjax.rstudio.com/latest/MathJax.js?config=TeX-MML-AM_CHTML' -V 'tit
le-slide-class=center, middle, inverse, title-slide' --standalone --section-divs --template
 /home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.5/xaringan/rmarkdown/templates/xaringan/
resources/default.html --no-highlight --include-in-header /tmp/RtmppuN4Ws/rmarkdown-strc3485
57c67.html --metadata pagetitle=20190212-Longitudinal_CI_Tutorial_X.utf8.md --include-before
-body /tmp/RtmppuN4Ws/xaringanc32ee0222b.md --include-after-body /tmp/RtmppuN4Ws/xaringanc32
14d3c48.js --variable math=true

Output created: 20190212-Longitudinal_CI_Tutorial_X.html
>
```

Environment | History | Connections

Import Dataset | List

Global Environment

Environment is empty

Files | Plots | Packages | Help | Viewer

New Folder | Upload | Delete | Rename | More

Cloud > project

Name	Size	Modified
..		
20190212-Longitudinal_CI_Tutorial_X.Rmd	19.8 KB	Feb 7, 2019, 1:41 PM
project.Rproj	205 B	Feb 7, 2019, 1:39 PM
SICS_logo_small.jpg	6.5 KB	Feb 6, 2019, 10:36 AM
20190212-Longitudinal_CI_Tutorial_X.html	12.6 KB	Feb 6, 2019, 10:34 AM
background-blur-darker.jpg	1.5 MB	Feb 6, 2019, 10:32 AM
complex.png	10 KB	Feb 6, 2019, 10:32 AM
not_complex.png	7.6 KB	Feb 6, 2019, 10:32 AM
pearl_cause.png	16.5 KB	Feb 6, 2019, 10:32 AM
time_varying.png	26.5 KB	Feb 6, 2019, 10:32 AM
IMSAbstractTemplate-JYH.pdf	47.1 KB	Feb 6, 2019, 10:30 AM

2. For fast typers: Read code. Duplicate in Console.

Singapore Institute for Clinical Sciences
A*STAR

WORKED EXAMPLES

CASE STUDY #1

Smoking cessation on weight gain (NHEFS)

Quitting smoking on body weight change (NHEFS)

Question of interest: What is the effect of quitting smoking (A) in a population of adult U.S. smokers (1971-1982) on risk of weight gain (Y)?

Singapore Institute for Clinical Sciences

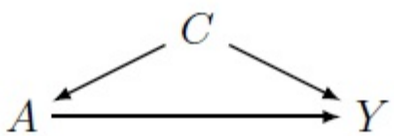
A*STAR

Quitting smoking on body weight change (NHEFS)

Question of interest: What is the effect of quitting smoking (A) in a population of adult U.S. smokers (1971-1982) on risk of weight gain (Y)?

Simplifying assumptions:

- ONLY factors confounding (C) the relationship (*i.e.* biasing distribution POs): *sex, age, race, income, marital status, schooling, physical activity, heart failure, high BP, asthma, diabetes, chronic bronchitis / emphysema*
- No informative missingness, and all C measured prior to A, time-invariant, and dichotomous such that causal relationships faithfully represented by:



Singapore Institute for Clinical Sciences

IMPORT DATA

Load NHEFS Data (NHANES 1 Epidemiologic Follow-Up Study)

```
aa <- read_csv(paste0(here(),"/nhefs.csv"))
# original sample size
nrow(aa)
```

```
## [1] 1746
```

Simplify data and create a weight gain indicator `delta` as outcome

```
a <- aa %>%
  select(seqn,qsmk,smkintensity82_71,smokeintensity,active,exercise,wt82_71,
         sbp,dbp,hbp,hf,ht,hbpmed,sex,age,hf,race,income,marital,school,
         asthma,bronch,diabetes) %>%
  mutate(hbp_71 = hbp) %>% na.omit()

a$delta <- as.numeric(a$wt82_71>0)

a <- a %>% select(delta, qsmk, sex, age, race, income, marital, school, active,
                  hf, hbpmed, asthma, bronch, smokeintensity, exercise,
                  diabetes, hbp_71)
```

WARNING: Complete case not recommended in general practice

Singapore Institute for Clinical Sciences
A*STAR

To reduce residual confounding, continuous or interval measures should not be casually dichotomized. However, we will do so only for the purposes of this exercise:

```
a$smokeintensity <- as.numeric(a$smokeintensity>median(a$smokeintensity))
a$age <- as.numeric(a$age>median(a$age))
a$exercise <- as.numeric(a$exercise>0)
a$income <- as.numeric(a$income>median(a$income))
a$marital <- as.numeric(a$marital>median(a$marital))
a$school <- as.numeric(a$school>median(a$school))
a$active <- as.numeric(a$active>0)
a$hbpmed <- as.numeric(a$hbpmed>0)
a$smokeintensity <- as.numeric(a$smokeintensity>median(a$smokeintensity))
a$exercise <- as.numeric(a$exercise>0)
a$diabetes <- as.numeric(a$diabetes>0)
a$hbp_71 <- as.numeric(a$hbp_71>0)
```

DESCRIBE DATA

Describe population and covariates, *e.g.*:

```
aa %>% ggplot() + geom_histogram(aes(age))
```

Describe basic exposure-outcome relationship:

```
a %>% group_by(qsmk) %>%
  summarize(gain = sum(delta),
            no_gain = sum(!delta),
            mu = round((gain)/(gain + no_gain), 2))
```

```
## # A tibble: 2 x 4
##    qsmk  gain no_gain    mu
##   <dbl> <dbl>   <int> <dbl>
## 1     0   724     390  0.65
## 2     1   268      94  0.74
```

Singapore Institute for Clinical Sciences

A*STAR

TRANSLATE DAG TO MODELS

Causal quantity of interest:

$$E(Y^{a=1} - Y^{a=0})$$

DAG gives us:

Variable	Modelled by
Y	$E(Y \mid A, C) = \alpha_0 + \alpha_1 A + \alpha_2 \mathbf{C}$
A	$P(A = 1 \mid C) = \text{expit}(\beta_0 + \beta_1 C)$

**expit(a) = 1/[1+exp(-a)]*

Causal assumptions allow us to transform:

$$E(Y) = \sum_A \sum_C E(Y \mid A, C) P(A \mid C) P(C)$$

into:

$$E(Y^a) = \sum_C E(Y \mid A = a, C) P(C)$$

when assigning a value $A = a$.

Singapore Institute for Clinical Sciences

A*STAR

FIT MODELS

Fit models for A and Y conditional on covariates:

```
model_A <- glm(qsmk ~ sex+age+race+income+marital+school+active+hf+hbpmed+asthma+
                 bronch+smokeintensity+exercise+diabetes+hbp_71,
               data=a,family=binomial("logit"))
summary(model_A)

model_Y <- glm(delta ~ qsmk+sex+age+race+income+marital+school+active+hf+hbpmed+
                 asthma+bronch+smokeintensity+exercise+diabetes+hbp_71,
               data=a,family=binomial("logit"))
summary(model_Y)
```

Note: Can survey some of the estimates for face validity. But should avoid over-interpretation, due to the joint conditioning ("Table 2 Fallacy").

Singapore Institute for Clinical Sciences
A*STAR

RESAMPLE FOR SIMULATION / DIAGNOSTICS

Monte Carlo resampling (with replacement). Usually, *N* * some factor (1 - 1000).

```
# resample data
index <- sample(1:nrow(a),size=1e4,replace=T)
MC <- a[index,]
nrow(MC)
```

```
## [1] 10000
```

```
MC$qsmk<-NULL
# predict exposure
pA <- predict(model_A,newdata=MC,type="response")
```

The variable `pA` is the predicted exposure under model.
Predicted prevalence of exposure: 0.2452247 versus observed prevalence: 0.2452575.

Singapore Institute for Clinical Sciences
A*STAR

SIMULATE DISCRETE EXPOSURE STATES

Transform predicted variable (0 < `pA` < 1) to binary exposure status (`qA`) by comparing to a uniform random value:

```
u <- runif(1e4)
qA <- as.numeric(pA>u)
head(qA)
```

```
## [1] 0 0 0 0 1 1
```

```
mean(qA)
```

```
## [1] 0.2501
```

```
mean(a$qsmk)
```

```
## [1] 0.2452575
```

Singapore Institute for Clinical Sciences

SIMULATE OUTCOMES UNDER BASELINE CONDITIONS

Taking new simulated exposure qA status, we simulate the outcome under observed conditions:

```
pY <- predict(model_Y,newdata=data.frame(MC,qsmk=qA),type="response")

mean(pY)
```

```
## [1] 0.6733915
```

```
mean(a$delta)
```

```
## [1] 0.6720867
```

Note: While this is neither necessary nor sufficient to demonstrate a valid model, deviations from the observed outcome can indicate modelling or data errors.

Singapore Institute for Clinical Sciences
A*STAR

EFFECT OF SMOKING CESSATION

We can now estimate the effect of smoking on weight gain risk by:

1. Setting `qsmk` = 1, predict the probability of weight gain, taking the mean
2. Repeat with `qsmk` set to 0
3. Taking the difference (and ratio)

```
pY_1 <- predict(model_Y,newdata=data.frame(MC,qsmk=1),type="response")
mY_1<-mean(pY_1)

pY_0 <- predict(model_Y,newdata=data.frame(MC,qsmk=0),type="response")
mY_0<-mean(pY_0)

(RD <- round((mY_1 - mY_0)*100,2))
```

```
## [1] 11.39
```

```
(RR <- round(mY_1 / mY_0,2))
```

```
## [1] 1.18
```

Singapore Institute for Clinical Sciences
A*STAR

INTERPRETATION

Can we interpret RD = 11.39 more cases of weight gain per 100 persons (or RR = 1.18 times risk) causally?

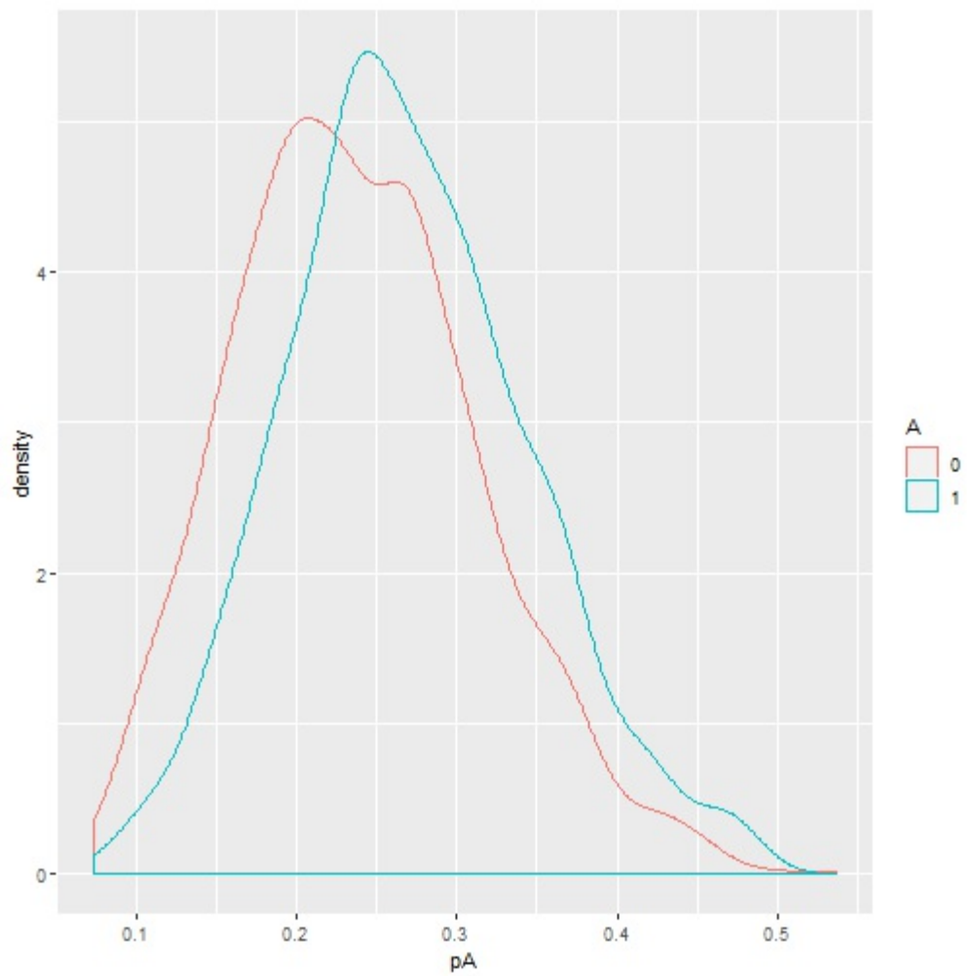
Evaluate causal assumptions:

- *Consistency*: Will different ways to quit smoking lead to different effects?
- Correct model specification and *exchangeability*: Did we include all the possible confounders? Residual confounding? Interaction?
- *Interference* (NEW): If a given person quit smoking, will it affect the outcome of another person in the study?
- *Positivity*: Are there are exposed and unexposed individuals in each confounder level?

POSITIVITY CHECK

```
prop <- model_A$fitted.values
propD <- data.frame(
  A=as.factor(a$qsmk),pA=prop)

ggplot(propD, aes(x=pA,color=A)) +
  geom_density()
```



Propensity score overlap:

Because of the number of confounding variables, we cannot use 2×2 tables.

Instead we can examine predicted probability of exposure.

Reasonable overlap between the two groups and likely no rare confounder combinations.

Singapore Institute for Clinical Sciences
A*STAR

CI BY BOOTSTRAP

Resample (with replacement) original data re-fitting outcome model; MC to compute contrast.

Repeat x 100 saving contrasts.

```
res <- NULL
for(i in 1:100){
  index <- sample(1:nrow(a),nrow(a),replace=T)
  boot_dat <- a[index,]
  model_Y <- glm(delta ~ qsmk+sex+age+race+income+marital+school+active+hf+hbpmed+asthma+bronch+
                 smokeintensity+exercise+diabetes+hbp_71,data=boot_dat,family=binomial("logit"))

  index <- sample(1:nrow(a),size=1e4,replace=T)
  MC <- boot_dat[index,]
  MC$qsmk<-NULL

  mY_1 <- mean(predict(model_Y,newdata=data.frame(MC,qsmk=1),type="response"))
  mY_0 <- mean(predict(model_Y,newdata=data.frame(MC,qsmk=0),type="response"))

  RD <- (mY_1 - mY_0)*100
  logRR <- log(mY_1 / mY_0)

  res <- rbind(res,cbind(RD,logRR))
}
```

Singapore Institute for Clinical Sciences

A*STAR

CI BY BOOTSTRAP

Take take SD across bootstraps as SE; calculate Wald confidence limits:

```
head(res)
```

```
##             RD      logRR
## [1,] 11.348235 0.16168739
## [2,]  9.082945 0.13487830
## [3,]  5.734421 0.08472272
## [4,] 12.565192 0.17826077
## [5,] 10.325501 0.14892598
## [6,] 11.409032 0.16151644
```

```
res_sd <- apply(res,2,sd)

lclRD <- RD - 1.96*res_sd[1]
uclRD <- RD + 1.96*res_sd[1]

lclRR <- exp(log(RR) - 1.96*res_sd[2])
uclRR <- exp(log(RR) + 1.96*res_sd[2])
```

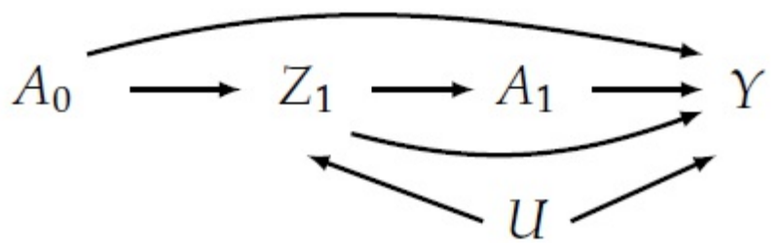
This bootstrap estimator yields 95% CIs of [2.51, 11.37] for the risk difference, and [1.11, 1.25] for the risk ratio.

Singapore Institute for Clinical Sciences
A*STAR

CASE STUDY 2

ART treatment on CD4 count

ART treatment on CD4 count



Working DAG

Where:

A_0 = ART treatment at $t = 0$
Z_1 = HIV viral load >200 copies/mL just before $t = 1$
A_1 = ART treatment at $t = 1$
Y = CD4 count (cells / mm^3)
U = unmeasured health status

Observed data on N = 100,000 patients:

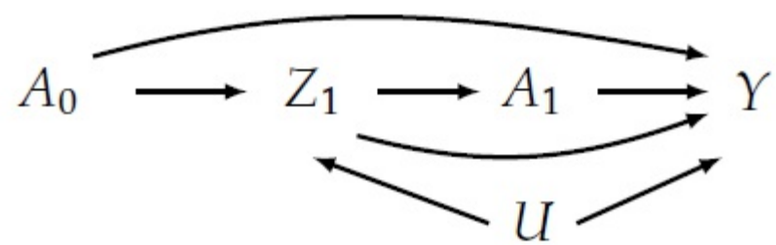
A_0	Z_1	A_1	Y	N
0	0	0	87.29	20,927
0	0	1	112.11	9,378
0	1	0	119.65	6,065
0	1	1	144.84	13,630
1	0	0	105.28	13,478
1	0	1	130.18	6,079
1	1	0	137.72	9,390
1	1	1	162.83	21,053

(Hypothetical cohort study)

Singapore Institute for Clinical Sciences
A*STAR

Question and statistical model

What is the effect of $\bar{A}_1 = \{1, 1\}$ on CD4 count (Y)?



Variable	Model
Y	$E(Y \mid A_1, Z_1, A_0) = \alpha_0 + \alpha_1 A_1 + \alpha_2 Z_1 + \alpha_3 A_0$
A_1	$P(A_1 \mid Z_1) = \text{expit}(\beta_0 + \beta_1 Z_1)$
Z_1	$P(Z_1 \mid A_0) = \text{expit}(\gamma_0 + \gamma_1 A_0)$
A_0	$P(A_0) = \text{expit}(\theta_0)$

Singapore Institute for Clinical Sciences
A*STAR

Statstical model to g-formula

$$E(Y) = \sum_{A_1} \sum_{Z_1} \sum_{A_0} E(Y \mid A_1, Z_1, A_0) Pr(A_1 \mid Z_1) P(Z_1 \mid A_0) Pr(A_0)$$

Under causal assumptions, we can:

- replace all A_0 and A_1 with a_0 and a_1 (respectively),
- remove models for A_0 and A_1 (because they are assigned),
- and equate observed conditional outcome with their PO

$$E(Y^{a_0, a_1}) = \sum_{Z_1} E(Y \mid A_1 = a_1, Z_1, A_0 = a_0) Pr(Z_1 \mid A_0 = a_0)$$

Then the objective is to estimate the conditional means and probabilities.

Singapore Institute for Clinical Sciences

A*STAR

INPUT DATA

```
a0<-c(0,0,0,0,1,1,1,1)
z1<-c(0,0,1,1,0,0,1,1)
a1<-c(0,1,0,1,0,1,0,1)
y<-c(87.29,112.11,119.65,144.84,105.28,130.18,137.72,162.83)
N<-c(20927,9378,6065,13630,13478,6079,9390,21053)
#N<-c(209271,93779,60654,136293,134781,60789,93903,210530)
D<-NULL
for(i in 1:8){
  d<-data.frame(cbind(rep(a0[i],N[i]),rep(z1[i],N[i]),
                      rep(a1[i],N[i]),rep(y[i],N[i])))
  D<-rbind(D,d)
}
names(D)<-c("a0","z1","a1","y")
```

Sample size: 100000

Initially treated: 0.5

Mean CD4 count: 125.0951071

Singapore Institute for Clinical Sciences

A*STAR

We fit each of the statistical models implied by our DAG:

```
mA0<-glm(a0~1,data=D,family=binomial("logit"))
mZ1<-glm(z1~a0,data=D,family=binomial("logit"))
mA1<-glm(a1~z1,data=D,family=binomial("logit"))
mY<-glm(y~a1+z1+a0,data=D,family=gaussian("identity"))
```

Predict each variable in the correct order:

```
# A0 (50% as observed)
pA0<-predict(mA0,type="response")
# Z1 <- A0
pZ1<-predict(mZ1,newdata=data.frame(a0=pA0),type="response")
# A1 <- ZI
pA1<-predict(mA1,newdata=data.frame(z1=pZ1),type="response")
# Y <- A0, Z1, and A1
pY<-predict(mY,newdata=data.frame(a0=pA0,z1=pZ1,a1=pA1),type="response")
```

Predicted initially treated: 0.5

Simulate mean CD4 count under "natural course": 125.102381

Sample mean CD4 count: 125.0951071

ESTIMATE TARGET CONTRAST

Setting A{1,1}:

```
pZ_1<-predict(mZ1,newdata=data.frame(a0=1),type="response")
pY_1<-predict(mY,newdata=data.frame(a0=1,z1=pZ_1,a1=1),type="response")
mY_1<-mean(pY_1)
```

Mean CD4 count until always treat: 150.0640176

Setting A{0,0}:

```
pZ_0<-predict(mZ1,newdata=data.frame(a0=0),type="response")
pY_0<-predict(mY,newdata=data.frame(a0=0,z1=pZ_0,a1=0),type="response")
mY_0<-mean(pY_0)
```

Mean CD4 count until always treat: 100.0562062

Mean difference = 50 cells/mL (+25 cell/mL per time-point).

(Causal if assumptions hold to allow interpretation of g-formula.)

How about simple regression?

Effect of $A_1 \mid Z_1$ (true effect = +25 cells/ML):

```
round(coef(glm(y~a1+z1,data=D,family=gaussian("identity"))),1)
```

```
## (Intercept)          a1          z1
##        94.3        25.0        36.4
```

Effect of A_0 (true effect = +25 cells/ML):

```
round(coef(glm(y~a0,data=D,family=gaussian("identity"))),1)
```

```
## (Intercept)          a0
##       111.6        27.1
```

Overestimates the true effect by 2.1 cells/mL, because it includes both the direct effect ($A_0 \rightarrow Y$) ***plus*** indirect effect ($A_0 \rightarrow A_1 \rightarrow Y$).

Summing to get the total effect of A{1,1} would give incorrect estimate!

Singapore Institute for Clinical Sciences
A*STAR

CASE STUDY 3

Aspirin on live birth (EAGeR Trial)

Estimating per-protocol effects

ITT estimates effect of treatment randomization:

- may not be good estimate of drug efficacy with non-compliance

Standard regression approaches to per-protocol effect biased:

- non-compliance caused by treatment, related to POs
- stratification by compliance disrupts randomization
- may be temporary non-compliance

EAGeR Study

- *Effect of Aspirin in Gestation and Reproduction* (EAGeR)
- Multicenter, block randomized, double-blind RCT
- ASA versus placebo on live birth
- Women (18-40 y/o) with 1-2 pregnancy loss
- 81 mg ASA + folic acid / day (N = 614)
- folic acid only (N = 614)

Singapore Institute for Clinical Sciences
A*STAR

Other parameters

- Allowed up to 6 cycles to conceive
- Treatment up to 36 weeks gestation
- Baseline: age, BMI, income, race, education, marital status, employment, study site
- Time-varying: compliance, vaginal bleeding, GI symptoms, TTP
- Outcomes: live birth, pregnancy loss, loss-to-follow-up

Question of interest:
What is the effect of continual treatment by ASA on live birth among target population if compliance was maintained through end of follow-up (by birth, pregnancy loss, or withdrawal)?

READ DATA AND DESCRIBE

```
aspirin<-read.table("aspirin2.txt",header=T,sep="\t")

aspirin %>% names()
```

```
##  [1] "id"            "study_month"    "eligibility"    "age"
##  [5] "income"        "education"      "white"          "marital"
##  [9] "employed"      "BMI"            "compliance"     "treatment"
## [13] "bleeding"      "gastro"         "conceived"      "efuwp"
## [17] "pregnancy_loss" "live_birth"    "last"           "site"
```

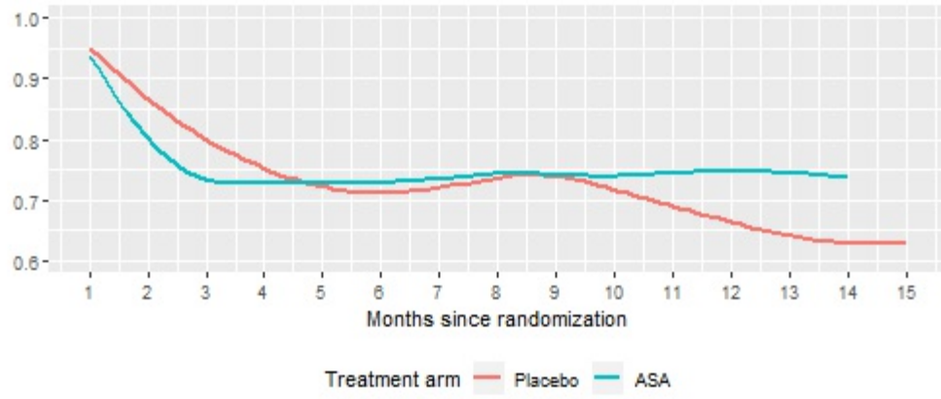
```
aspirin %>% filter(last == 1) %>% group_by(treatment) %>%
  summarize(Y_Birth = mean(live_birth), D_Loss = mean(pregnancy_loss),
            S_Censor = mean(efuwp))
```

```
## # A tibble: 2 x 4
##   treatment Y_Birth D_Loss S_Censor
##       <int>   <dbl>  <dbl>    <dbl>
## 1         0   0.485  0.194    0.303
## 2         1   0.549  0.192    0.259
```

Singapore Institute for Clinical Sciences
A*STAR

Compliance?

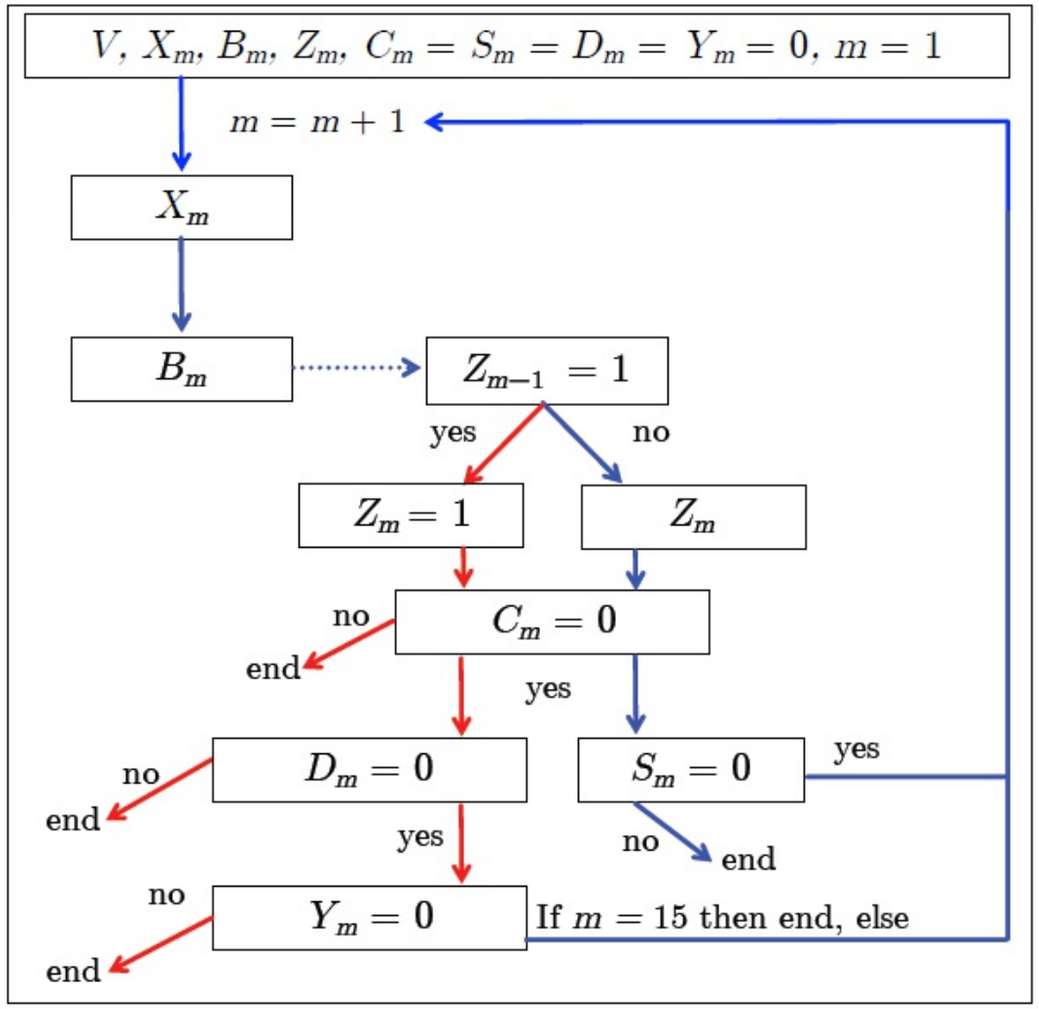
```
aspirin %>% mutate(tx = factor(treatment, levels = c(0,1), labels = c("Placebo", "ASA"))) %>%
  ggplot() + geom_smooth(aes(x = study_month, y = compliance, color = tx), se = F) +
  scale_y_continuous(limits = c(0,1)) + scale_x_continuous(breaks = c(1:15)) +
  labs(y = "", x = "Months since randomization", color = "Treatment arm") + theme(legend.position = "bottom") + coo
```



Singapore Institute for Clinical Sciences
A*STAR

Presumed causal ordering (at each time point):

Order	Variable	Description
8	Y	Live birth
7	D	pregnancy loss
6	S	No pregnancy
5	C	Withdrawal
4	Z	Conception
3	X	Compliance
2	N	GI Symptoms
1	B	Bleeding



Singapore Institute for Clinical Sciences
A*STAR

FIT MODELS BASED ON ORDERING

Order	Variable	Description
8	Y	Live birth
7	D	pregnancy loss
6	S	No pregnancy
5	C	Withdrawal
4	Z	Conception
3	X	Compliance
2	N	GI Symptoms
1	B	Bleeding

Singapore Institute for Clinical Sciences

FIT MODELS BASED ON ORDERING

Order	Variable	Description
8	Y	Live birth
7	D	pregnancy loss
6	S	No pregnancy
5	C	Withdrawal
4	Z	Conception
3	X	Compliance
2	N	GI Symptoms
1	B	Bleeding

e.g. Compliance `X` at each month k:

```
fitX<-glm(X~V1+V2+V3+V4+V5+V6+V7+
ns(V8,df=3)+V9+V10+V11+V12+V13+
ns(V14,df=3)+Xl+Xl1+B+Bl+Bl1+N+Nl+
Nl1+Z+Zl+Zl1+ns(jj,df=3),
family=binomial,data=boot,subset=R==k)
```

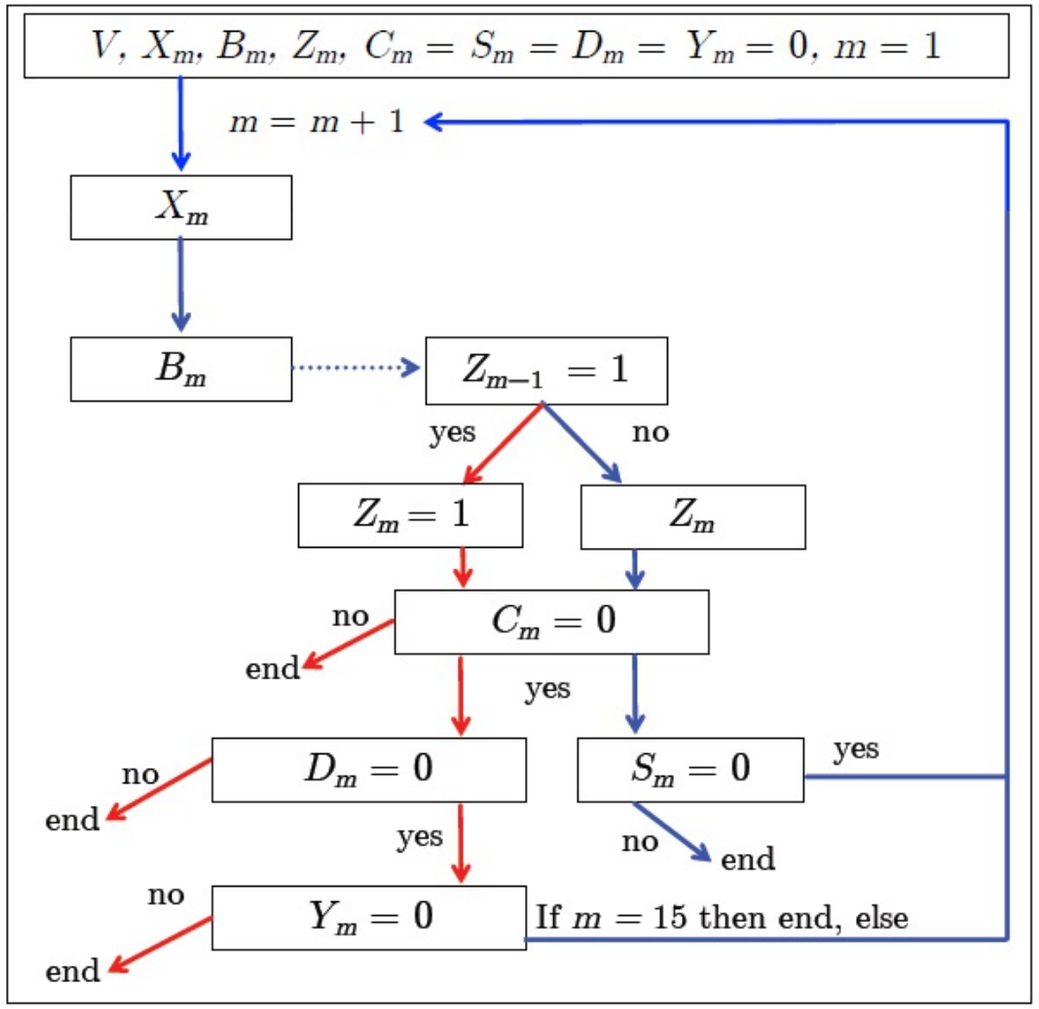
Modelled by natural cublic splines:
`V8` = maternal age @ baseline
`V14`= maternal BMI @ baseline
`jj` = months since enrollment (mean-centered)

Lag - two previous values: `Xl`, `Xl1`, `Bl`, `Bl1`, `Nl`, `Nl1`, `Zl`, `Zl1`

(Note: Can't be influenced by later events.)

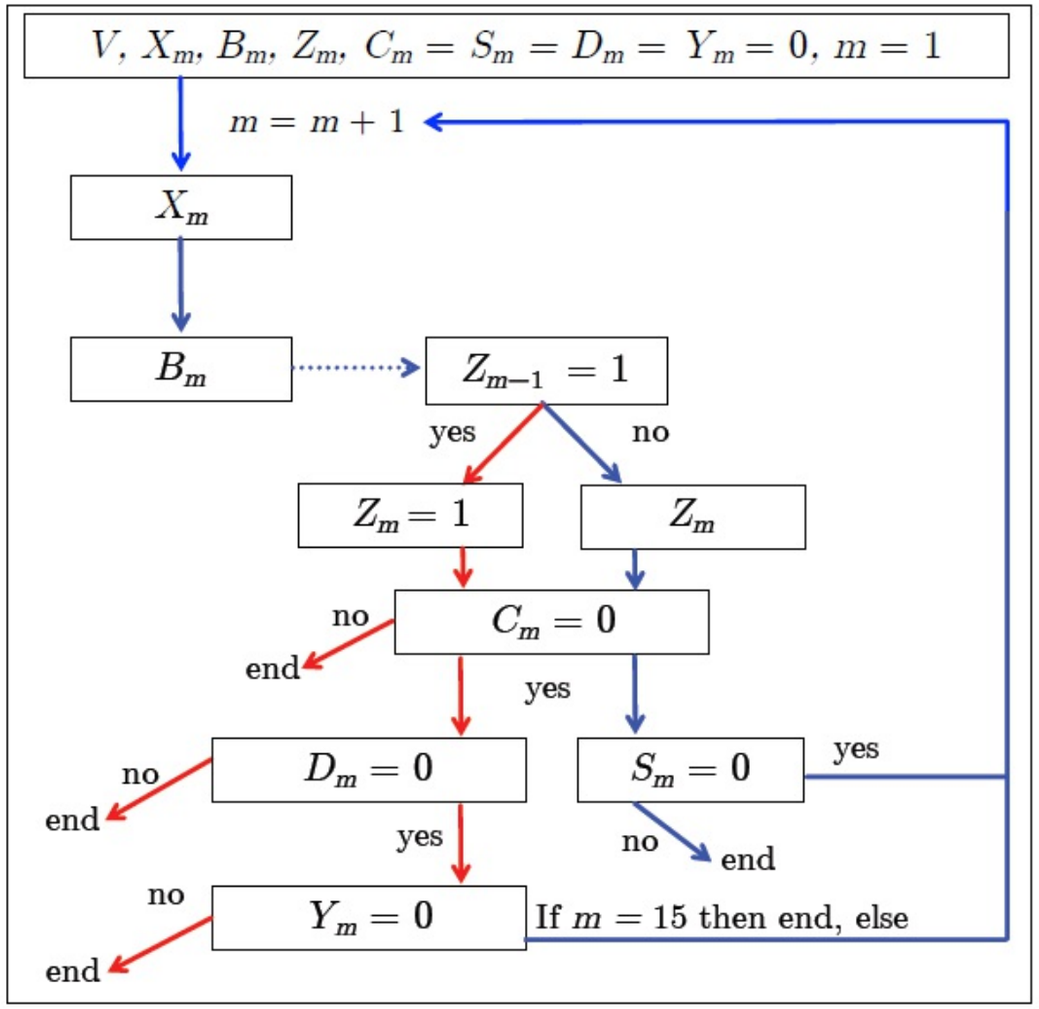
Singapore Institute for Clinical Sciences
A*STAR

Resample and iterate based on flowchart



Singapore Institute for Clinical Sciences

Resample and iterate based on flowchart



Simulation decisions based on trial characteristics:

- set live birth (Y) to 0 for all months < 8
- set pregnancy loss (D) to 0 for months < 2
- censor (S) at 6 months if no conception
- follow to 15 months

Singapore Institute for Clinical Sciences

RESULTS

MC resampled 500 × N (500 × 1228 = 614,000)

Simulate "natural course": g-formula: **0.5381526** (observed: **0.517101**)

```
mean(bn$Y,na.rm=T); mean(a2[a2$last==1,]$Y)
```

Per-protocol estimate, ASA: g-formula: **0.511022** (observed, ITT: **0.5488599**)

```
mean(b1$Y,na.rm=T); mean(a2[a2$last==1&R==1,]$Y)
```

Per-protocol estimate, placebo: g-formula: **0.448** (observed, ITT: **0.485342**)

```
mean(b0$Y,na.rm=T); mean(a2[a2$last==1&R==0,]$Y)
```

Per-protocol effect estimate: **+6.3% live birth**

(to run: pgf_annotated.R)

Singapore Institute for Clinical Sciences
A*STAR

Review: Estimation Steps

- Draw a DAG representing causal structure
- Model temporal / structural relationships (Q-model) implied by DAG
- Sample baseline values simulate POs using model
- Repeat setting exposures to levels reflecting desired contrast
- Bootstrap statistical model to obtain SE

Other considerations, methods, Q&A

Other estimation considerations

Other estimation considerations

1. *Missing data*
    - Single imputation within each bootstrap

Singapore Institute for Clinical Sciences

Other estimation considerations

1. *Missing data*
    - Single imputation within each bootstrap
2. *No interference*
    - Defining PO based on "distance" to other units

Singapore Institute for Clinical Sciences
A*STAR

Other estimation considerations

1. *Missing data*
    - Single imputation within each bootstrap
2. *No interference*
    - Defining PO based on "distance" to other units
3. *Test sensitivity to causal ordering*
    - Re-estimate with different Q-models

Singapore Institute for Clinical Sciences
A*STAR

Other estimation considerations

1. *Missing data*
    - Single imputation within each bootstrap
2. *No interference*
    - Defining PO based on "distance" to other units
3. *Test sensitivity to causal ordering*
    - Re-estimate with different Q-models
4. *Test sensitivity to simulation error*
    - Try out larger MC samples

Singapore Institute for Clinical Sciences
A*STAR

Other estimation considerations

1. *Missing data*
    - Single imputation within each bootstrap
2. *No interference*
    - Defining PO based on "distance" to other units
3. *Test sensitivity to causal ordering*
    - Re-estimate with different Q-models
4. *Test sensitivity to simulation error*
    - Try out larger MC samples
5. *Natural course*
    - Visualize means, medians, distributions

Singapore Institute for Clinical Sciences
A*STAR

Related approaches

Inverse-Probability Weighted Marginal Structural Models

- standardization by propensity score weighting
- *e.g.* counterfactual means for a pseudo-population
- (rather than standardized to observed covariate distributions)
- generate IPW for each treatment time
- slightly easier to implement

Singapore Institute for Clinical Sciences
A*STAR

Related approaches

Inverse-Probability Weighted Marginal Structural Models

- standardization by propensity score weighting
- *e.g.* counterfactual means for a pseudo-population
- (rather than standardized to observed covariate distributions)
- generate IPW for each treatment time
- slightly easier to implement

Targeted Maximum Likelihood-Based Estimation (TMLE)

- a g-computation approach that adds exposure model to "target" estimate
- usually implements cross-validated library of algorithms (SuperLearner) to estimate the Q-model:
  - lasso, regularlized GLM,
  - K nearest neighbors,
  - support vector machine,
  - random forests, *etc.*

Singapore Institute for Clinical Sciences
A*STAR

Some Helpful Resources

- Daniel RM, De Stavola BL, Cousens SN. gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *The Stata Journal.* 2011: 11(4). 479-517.
- Hernán MA, Robins JM (2019). *Causal Inference.* Boca Raton: Chapman & Hall/CRC, forthcoming.
  - Full working draft accessible at Miguel's site
- Naimi AI, Cole SR, Kennedy EH. An Introduction to G Methods. *Int J Epi.* 2016.
- Datasets & examples from Ashley Naimi's SER shortcourses.
  - Accessible at Ashley's github repo

Parting thoughts

- Clear thinking about desire estimands are essential
- Draw the DAG that best represents prior knowledge
- Natural intervals may or may not be defined by data
- Careful consideration of causal assumptions is critical
- Dynamic treatment regimes (Friday!)
- Interpret carefully!
- Any questions / comments: Email me!

Singapore Institute for Clinical Sciences
A*STAR

Final Word

Judea Pearl @yudapearl · 20 Nov 2018

Contrary to expectations, the definition of "causal modeling" is fairly easy to articulate. To me, "causal model" is a set of assumptions about the data generating process, which cannot be expressed as properties of the joint distribution of observed variables. #Bookofwhy

5 26 100

Singapore Institute for Clinical Sciences

A*STAR

ADDITIONAL SLIDES

Hernán, Hsu, Healy (2018)

Description - Quantiative summary of features of the world

Prediction - Map input features onto output features

Causal Inference - Predict different worlds given change of certain features (counterfactual)

Data science is science's second chance to get causal inference right: A classification of data science tasks.
(arXiv:1804.10846)

Singapore Institute for Clinical Sciences
A*STAR