Directed Acyclic Graphs (DAGs)

Gentle introduction

Mabel Carabali

EBOH, McGill University

Updated: 2025-09-04

Objectives

1. Operationalize Directed Acyclic Graphs (DAGs)

2. Appreciate the insights into confounding and selection bias provided by DAGs

3. Examples to appreciate the importance of DAGs (and their encoded substantive knowledge) on the road to causal inference

Felix, qui potuit rerum cognoscere causa - Vigil (29BC)

"Fortunate is he, who is able to know the causes of things"

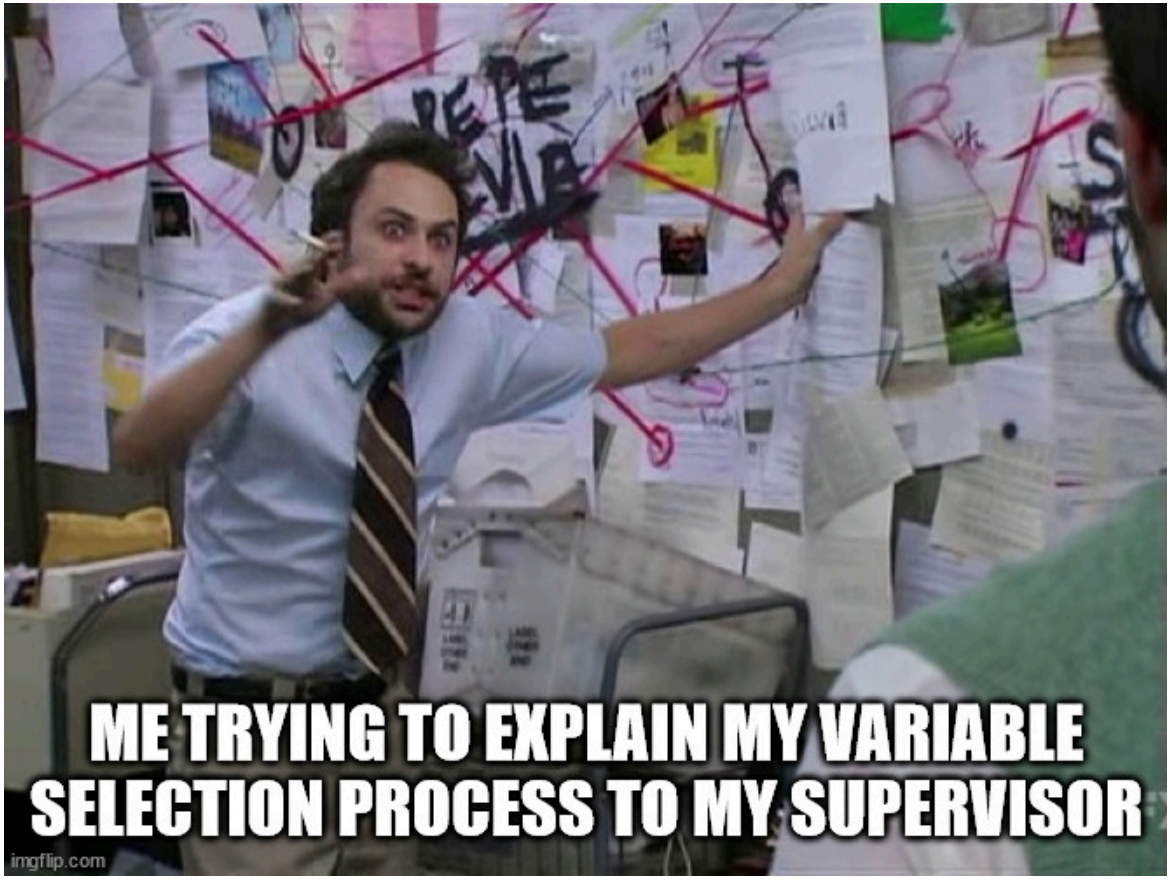
Included modified notes from Dr. Jay Brophy

Expected Competencies

- Knows what is a Directed Acyclic Graphs (DAGs)
- Knows basic considerations of cause-effect, causal inference.

Conventional statistics & causal inference

- “The object of statistical methods is the reduction of data” (Fisher 1922)
- Provides a parsimonious mathematical description of the joint distribution of observed variables
- Good statistical processes can describe the data
- But say nothing about the data generating process and **can’t answer causal questions**



From McGill: SPGH's and EBOSS Research Day: 2022 meme competition (item 12)

DAGs and causal inference

- **DAGs (AKA causal diagrams)** characterize causal structures compatible with the observations & assist in drawing logical conclusions about the statistical relations

- Help understanding: confounding, selection bias, covariate selection, over adjustment, instrumental variable analyses & avoid making errors about the statistical relations (much of this associated with Judea Pearl's work)

- **Potential Outcome (counterfactual)** framework provides another approach to causal inference, building on the work on RCTs from the 1920s by Fisher and Neyman (much of this associated with Donals Rubin's work)

These frameworks are largely complementary

Current statistical approach

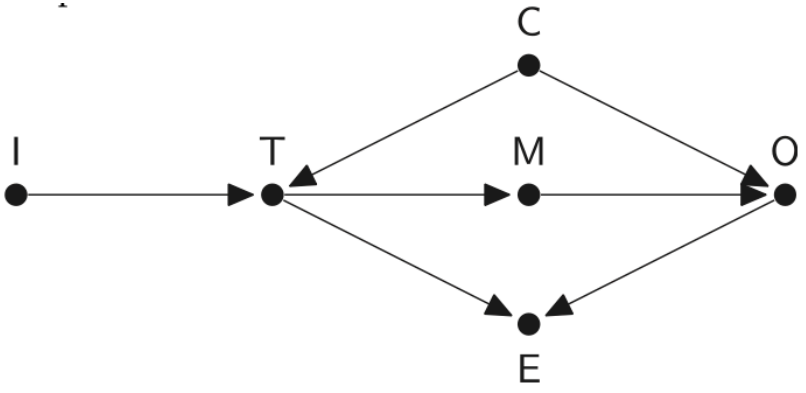
- Vague research question with some available relevant data
- “Let the data speak” (If the data are speaking to you...)
- Add all the variables and let multiple regression sort it out
- Regression models alone insufficient no distinction between causes, confounders, mediators and colliders
- Residual confounding, measurement error & missing data are often ignored
- Often model selection chosen via Akaike Information Criterion $(AIC = 2K - 2ln(logLikelihood))$ where K is the number of model parameters
- Provides the best predictive but not the causally correct model $\rightarrow$ **doesn't provide causal statements**

Canons of causal inference

- Every causal inference task must rely on judgmental, extra-data assumptions (or experiments)
- Ways of encoding those assumptions mathematically and test their implications exist
- DAGs encode qualitatively *a priori* subject matter knowledge
- Consideration of the causal model combined with data provides clarity in interpreting statistical coefficients and causal inferences

Assumption - free causal inference doesn't exist

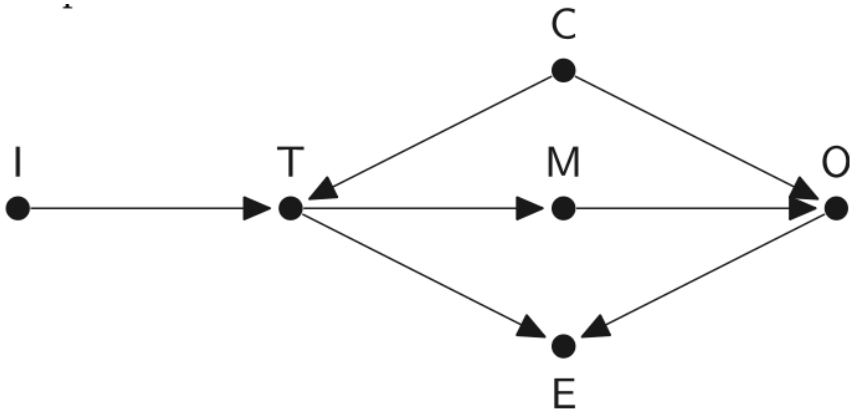
DAGs - Help identifying causal effects



1. **Treatment/intervention/exposure (T)**: the main cause.
2. **Outcome (O):** the main effect.
3. **Mediator (M):** caused by the treatment which in turn causes the outcome.
4. **Confounder (C):** common cause of the treatment and outcome.
5. **Collider (E):** common effect of any two variables on a backdoor path.*
6. **Instrument (I):** only causes the treatment (and not the outcome).

*Non-causal path from the treatment to the outcome.

DAGs - Help identifying causal effects

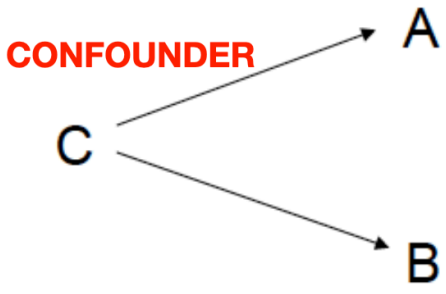


- **Variables** are depicted as **nodes** and connected by **arrows**
- Acyclic (the future can't predict the past)
- Missing lines strongest assumption, implies variable independence
- Include all common causes of any 2 variables & all variables involved in data generation (observed or unobserved)
- Contain both causal and non-causal pathways
- **Help identify causal effects by deriving testable implications of a causal model**

DAGs Between Two Variables

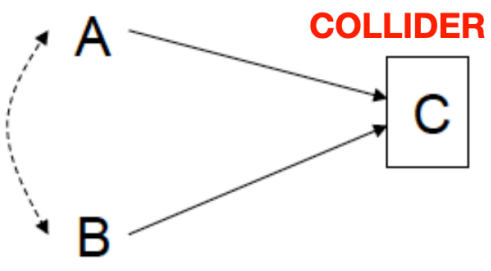
(1) Direct and indirect causation

$A \not\perp\!\!\!\perp B$ and $A \perp\!\!\!\perp B|C$



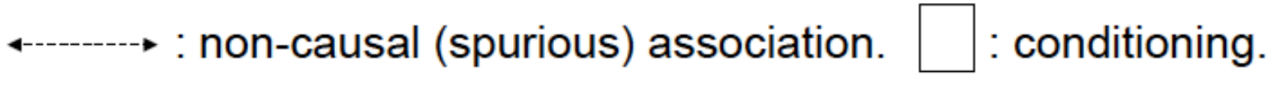
(2) Common cause confounding

$A \not\perp\!\!\!\perp B$ and $A \perp\!\!\!\perp B|C$

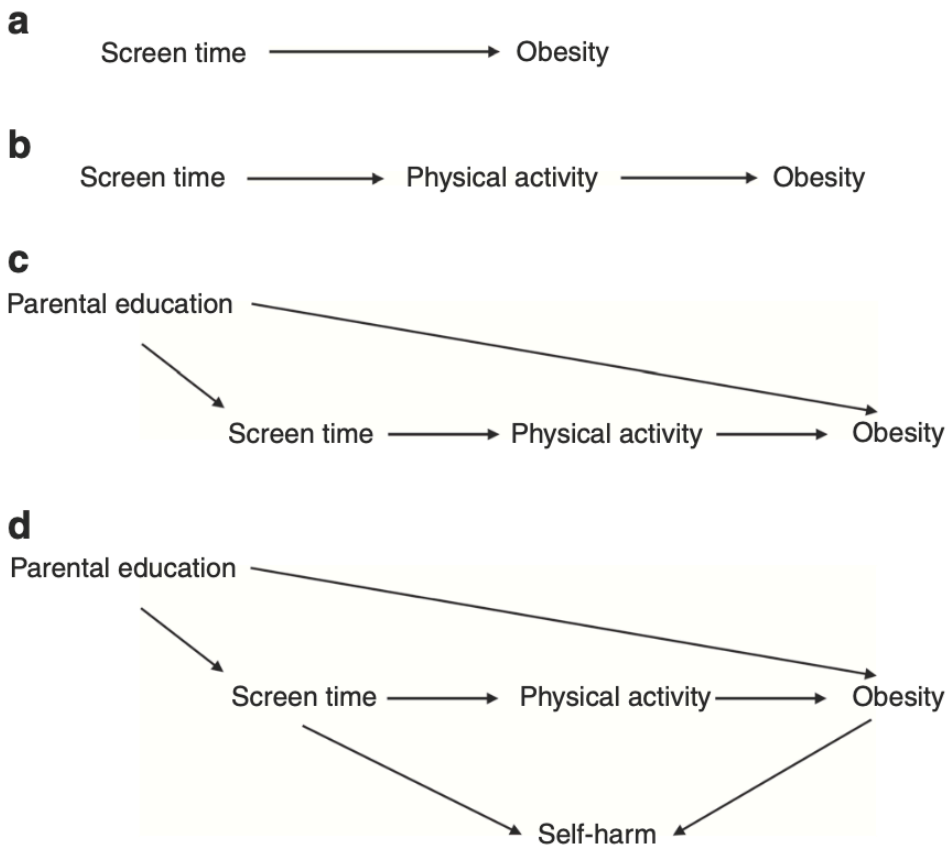


(3) Conditioning on a common effect ("collider"): Selection

$A \perp\!\!\!\perp B$ and $A \not\perp\!\!\!\perp B|C$

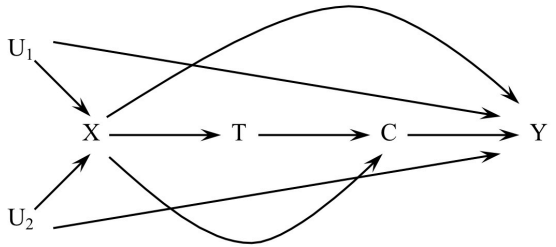


Directed acyclic graphs: a tool for causal studies in paediatrics



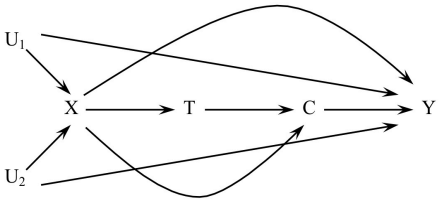
- a.Screen time (the exposure) **causes** obesity (the outcome).
- b. Screen time acts on obesity through the **mediator** of physical activity.
- c. Low parental education increases both screen time and obesity, and is therefore a **confounder**.
- d. Self-harm is a **collider** in the path from screen time to obesity.

More DAG terminology



- Path is a sequence of non-intersecting adjacent edges **X->T->C** or **U2->Y<-C<-T**
- Causal path a path in which all arrows point away from T to outcome Y; **T->C->Y**
- Total causal effect of a treatment on an outcome consists of all causal paths connecting them
- Non-causal path connecting T and Y with at least one arrow against flow of time **T<-X->Y**
- Descendants of a node: all nodes directly or indirectly caused by the node; **desc(T) = {C,Y}**
- Children of a node: all nodes directly caused by the node; **child(T) = {C}**
- Ancestors of a node: all nodes directly or indirectly causing the node; **an(T) = {X, U1, U2}**
- Collider variable along a path with 2 arrows pointing in **U->X<-U2**

More DAG terminology



- "Blocked" (d-separated) paths don't transmit associations
- "Unblocked" (d-connected) paths may transmit association
  Three blocking criteria
  - Conditioning on a non-collider blocks a path
  - Conditioning on a collider, or a descendent of a collider, unblocks a path
  - Not conditioning on a collider leaves a path "naturally" blocked

Implication:

- If X and Y are d-separated by Z along all paths in a DAG, then X is statistically independent of Y conditional on Z in every distribution compatible with the DAG
- If X and Y are not d-separated by Z along all paths in the DAG, then X and Y are dependent conditional on Z in at least one distribution compatible with the DAG

Identification

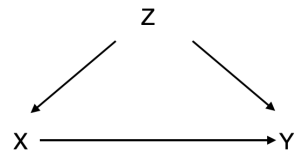
The causal effect of X on Y is said to be **"identified"** if it is possible, with ideal data (infinite sample size, perfect measurement), to purge all non-causal association from the observed association between X and Y such that only the causal association remains.

One way to interpret this with DAGs, is to note that the total causal effect of X on Y is identifiable if one can condition on ("adjust for") a set of variables {Z} that

1. Blocks all non-causal paths between X and Y,
2. Without blocking any causal paths between X and Y

Estimating causal effects

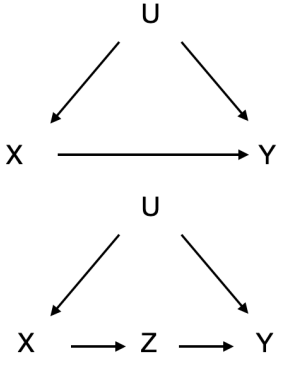
Backdoor criteria



Z is a sufficient set if:

1. No variable in Z is a descendant of X and
2. Every path between X and Y that contains an arrow into X is blocked by Z

Frontdoor criteria



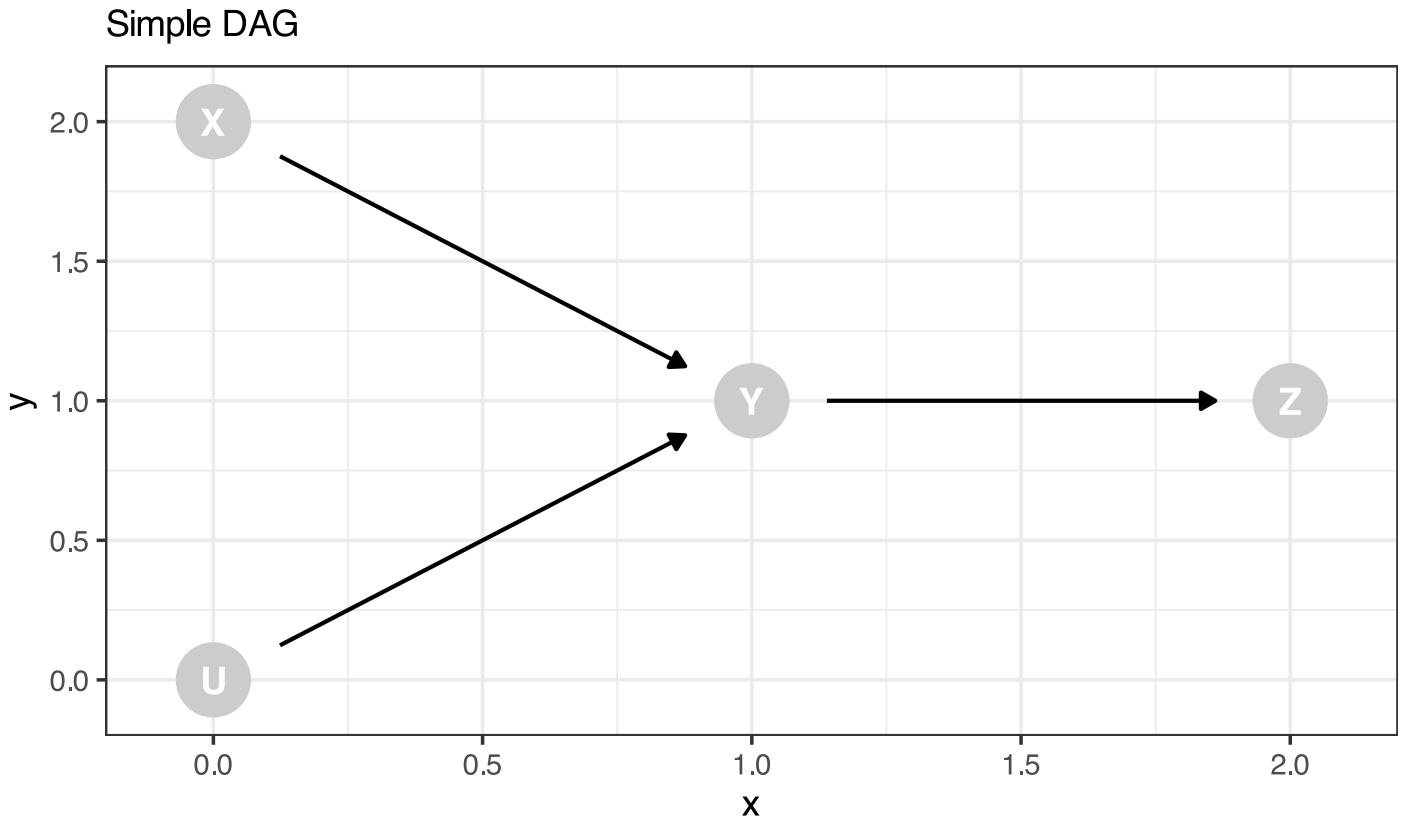
Z is a sufficient set if:

1. Z intercepts all directed paths from X to Y
2. No unblocked paths from X to Z
3. All backdoor paths from Z to Y are blocked by X

Pearl, Judea. Causal Inference in Statistics : A Primer, John Wiley & Sons, Incorporated, 2016. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/mcgill/detail.action?docID=7104473.

Simple DAG

What are the assumptions & statistical implications of this model?

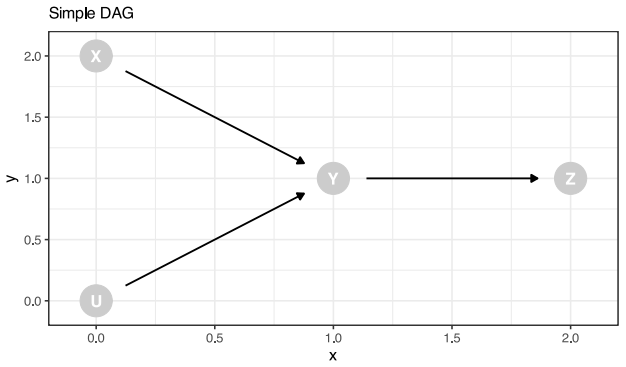


Would you believe at least 16 assumptions and statistical implications!

Simple DAG ? https://bookdown.org/jbrophy115/bookdown-clinepi/causal.html

Causal implications

1. X & U are direct causes of Y
2. Y is a direct cause of Z
3. X is an indirect cause of Z via Y
4. X is not a cause of U and U is not a cause of X
5. U is an indirect cause of Z via Y
6. No variable causes both X and Y OR U and Y



Statistical implications

1. X and Y are statistically dependent
2. U and Y are statistically dependent
3. Y and Z are statistically dependent
4. X and Z are statistically dependent
5. U and Z are statistically dependent
6. X and U are statistically independent
7. X and U are statistically dependent, conditional on Y
8. X and U are statistically dependent, conditional on Z
9. X and Z are statistically independent, conditional on Y
10. U and Z are statistically independent, conditional on Y

What does automated statistical software do?

- Let's consider the scenario where we assess systolic blood pressure (SBP) by *group*.
- Let's simulate data where SBP is a function of age but independent of racial group:

```
library(tidyverse)
n <- 200; age <- runif(n, 25, 65)
#sbp is not a function of group
sbp <- 99 + 0.1*age + exp(age/15) + rnorm(200, 0, sd = 5)
dat <- data.frame(age=age, sbp = sbp) %>%
  arrange(age) %>%
  mutate(group = case_when(
    age <40  ~ "0",
    age >= 40 ~ "1"))
dat$group <- as.factor(dat$group)
head(dat, 3)
```

```
##        age      sbp group
## 1 25.05423 107.1075     0
## 2 25.12015 107.1512     0
## 3 25.28153 105.8815     0
```

How would you analyze the data to assess this relationship?

Traditional (Standard) approach

```
##
## Call:
## lm(formula = sbp ~ group, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.954  -9.334  -0.971   7.098  41.427
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  111.369      1.631   68.28   <2e-16 ***
## group1        31.890      2.180   14.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.3 on 198 degrees of freedom
## Multiple R-squared:  0.5195,    Adjusted R-squared:  0.5171
## F-statistic: 214.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Spurious statistically difference in SBP by *group* is observed, yet data was generated with no group exposure effect.

Traditional (Standard) approach, OLS adjusting for *group*:

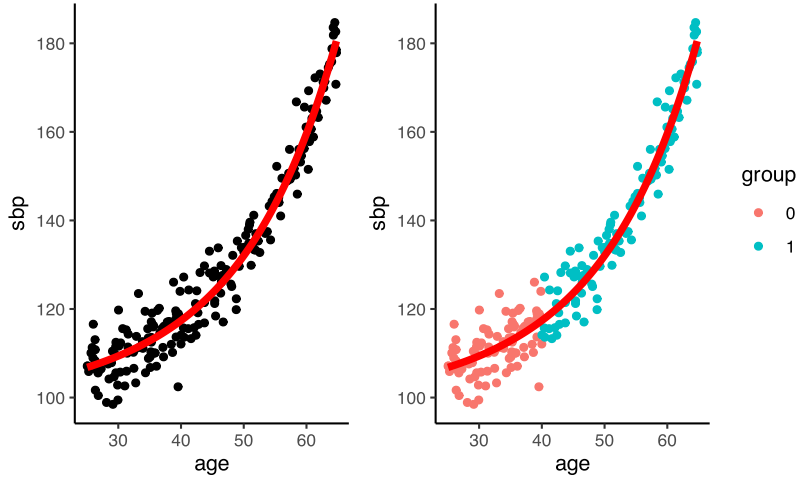
```
summary(lm(sbp ~ age + group, data=dat))
```

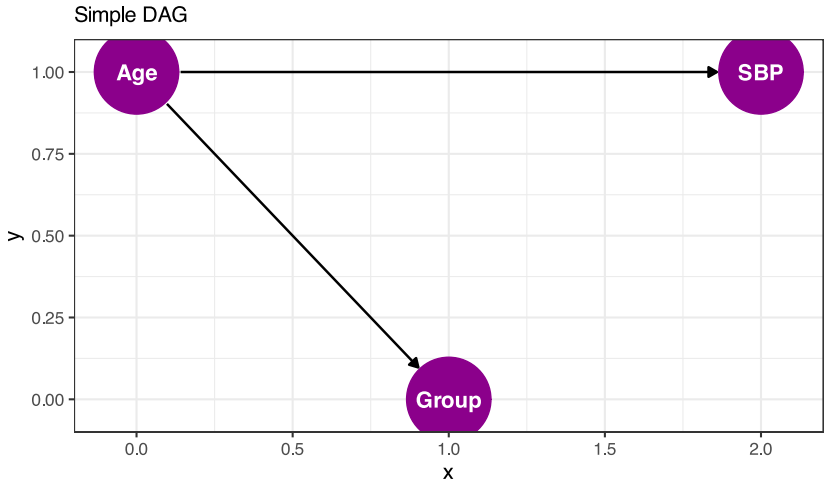
```
##
## Call:
## lm(formula = sbp ~ age + group, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.256  -5.000  -1.059   5.036  19.661
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.52879    2.76471  15.021  < 2e-16 ***
## age           2.12902    0.08095  26.302  < 2e-16 ***
## group1      -11.22161    1.93523  -5.799 2.63e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.222 on 197 degrees of freedom
## Multiple R-squared:  0.8935,    Adjusted R-squared:  0.8924
## F-statistic: 826.3 on 2 and 197 DF,  p-value: < 2.2e-16
```

Spurious statistically difference in SBP by *group* is observed, yet data was generated with no group exposure effect.

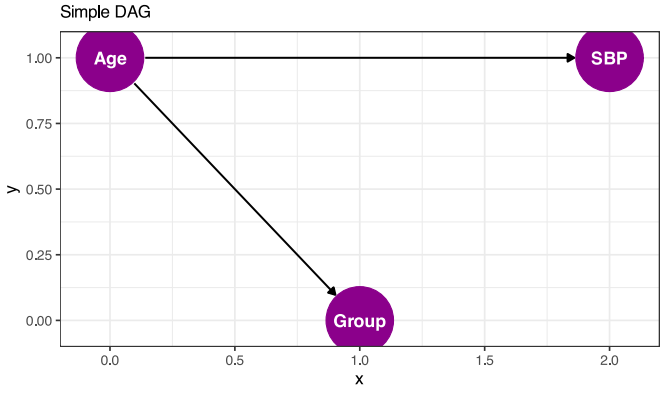
Visualizations and DAGs can help Plot the generated data

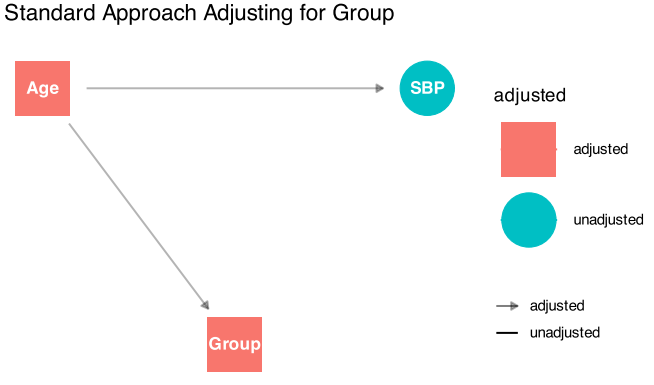
```
myfun <- function(age)  99 + 0.1*age + exp(age/15)
g1 <- ggplot(dat, aes(age, sbp)) + geom_point()  +
  stat_function(fun = myfun, color="red", size=1.8) +
  theme_classic()
g2 <- ggplot(dat, aes(age, sbp, color=group)) + geom_point()  +
  stat_function(fun = myfun, color="red", size= 1.8) +
  theme_classic()
```





Visualizations and DAGs can help





- While the DAG shows only one SBP causal pathway (Age $\longrightarrow$ SBP),
- The automatic software also includes the **spurious non-causal pathway:**

(Group $\longleftarrow$ Age $\longrightarrow$ SBP)

Would increasing the sample size help?

- Let's consider the scenario where we assess systolic blood pressure (SBP) by *group*.
- Let's simulate data where SBP is a function of age but independent of racial group:

```
library(tidyverse)
n <- 2000; age <- runif(n, 25, 65)
#sbp is not a function of group
sbp <- 99 + 0.1*age + exp(age/15) + rnorm(2000, 0, sd = 5)
dat1 <- data.frame(age=age, sbp = sbp) %>%
  arrange(age) %>%
  mutate(group = case_when(
    age <40  ~ "0",
    age >= 40 ~ "1"))
dat1$group <- as.factor(dat1$group)
head(dat1, 3)
```

```
##        age      sbp group
## 1 25.01826 111.9241     0
## 2 25.03487 103.1393     0
## 3 25.04956 109.9988     0
```

OLS without "adjustment"

```
L2mod2<- lm(sbp ~ group, data=dat1)
jtools::summ(L2mod2, model.info = F, model.fit = F, digits=3, confint=T)
```

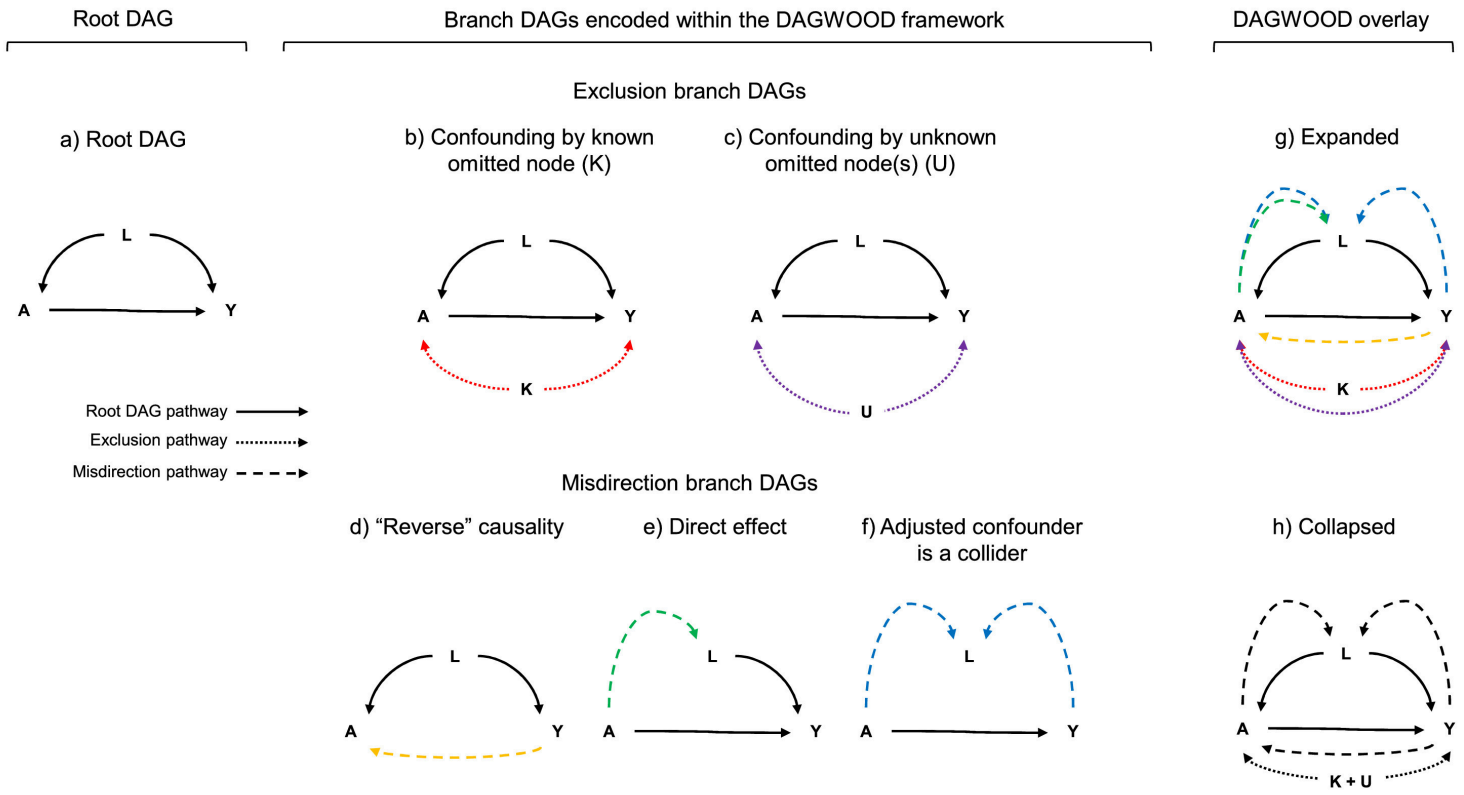
	Est.	2.5%	97.5%	t val.	p
(Intercept)	111.467	110.364	112.570	198.269	0.000
group1	29.736	28.348	31.124	42.016	0.000
Standard errors: OLS					

OLS adjusting for group

```
L2mod2a<-lm(sbp ~ age + group, data=dat1)
jtools::summ(L2mod2a, model.info = F, model.fit = F, digits=3, , confint=T)
```

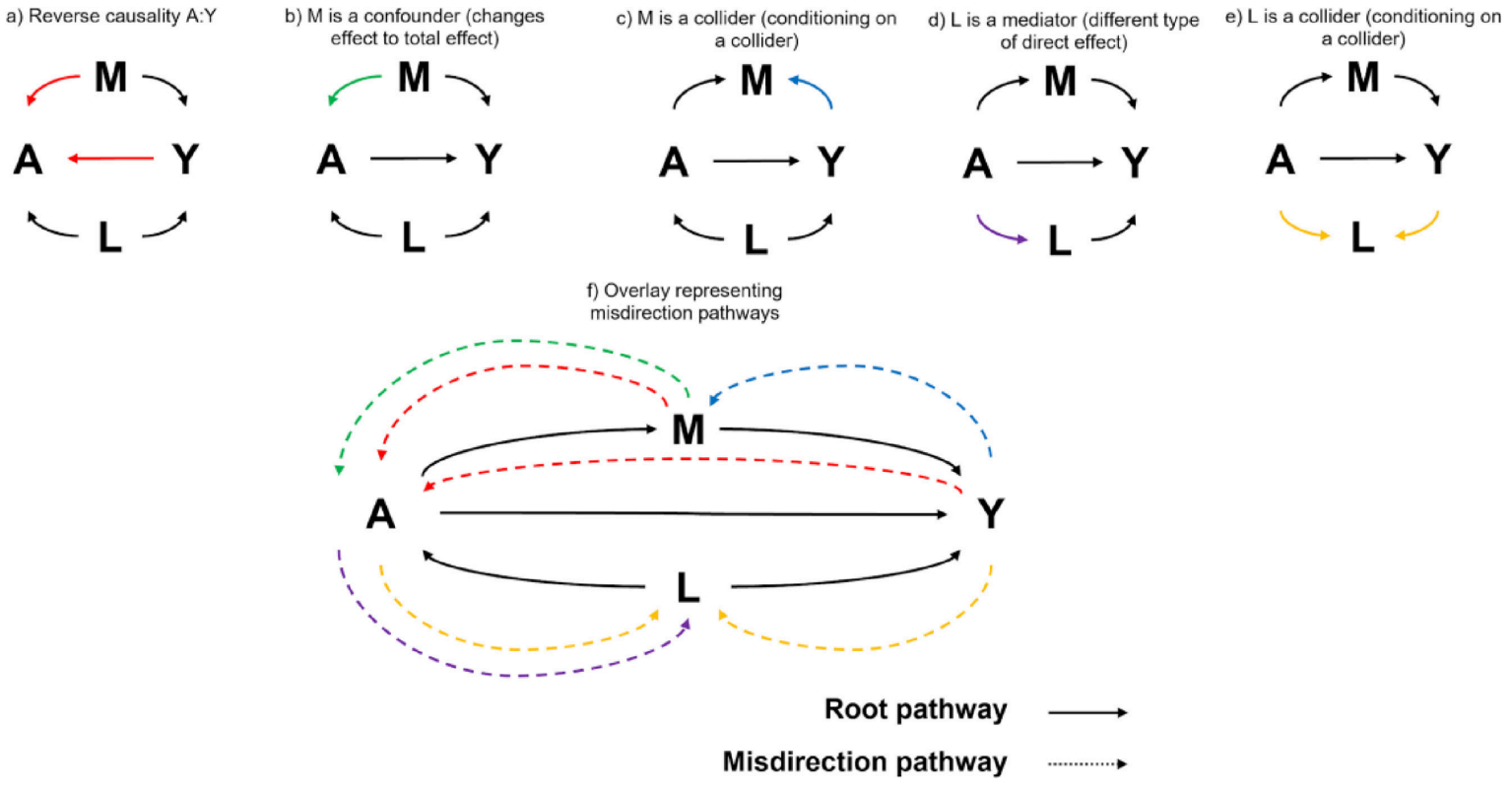
	Est.	2.5%	97.5%	t val.	p
(Intercept)	41.483	39.790	43.176	48.046	0.000
age	2.155	2.105	2.204	85.053	0.000
group1	-13.316	-14.500	-12.132	-22.052	0.000
Standard errors: OLS					

DAG With Omitted Objects Displayed (DAGWOOD): a framework for revealing causal assumptions in DAGs



Haber NA, Wood ME, Wieten S, Breskin A. DAG With Omitted Objects Displayed (DAGWOOD): a framework for revealing causal assumptions in DAGs. Ann Epidemiol. 2022 Apr;68:64-71. https://doi.org/10.1016/j.annepidem.2022.01.001

DAG With Omitted Objects Displayed (DAGWOOD): a framework for revealing causal assumptions in DAGs

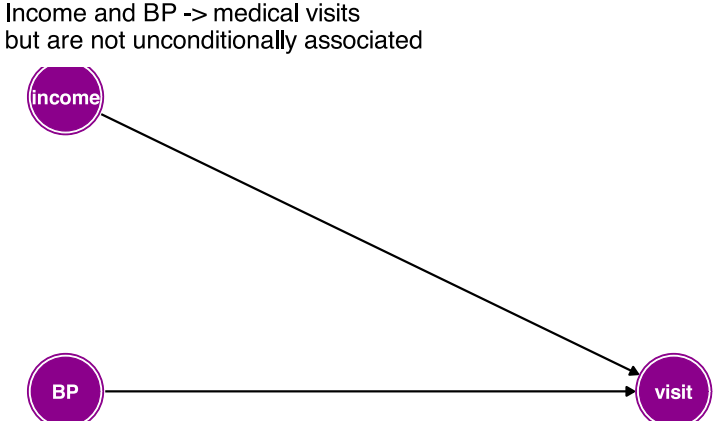


Haber NA, Wood ME, Wieten S, Breskin A. DAG With Omitted Objects Displayed (DAGWOOD): a framework for revealing causal assumptions in DAGs. Ann Epidemiol. 2022 Apr;68:64-71. https://doi.org/10.1016/j.annepidem.2022.01.001

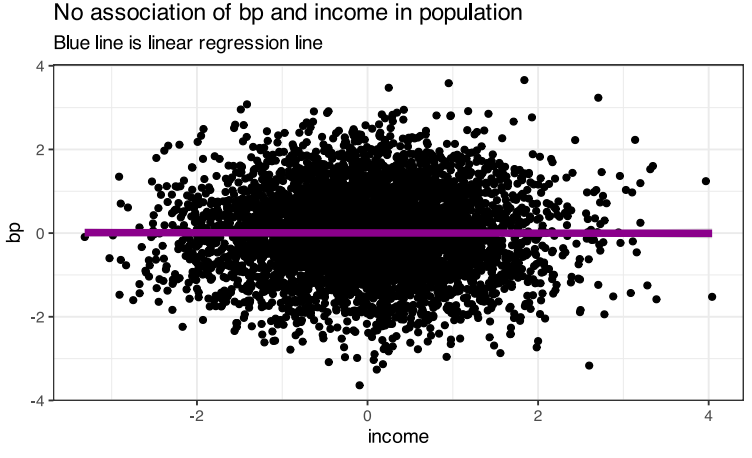
Understanding Selection Bias

Let's simulate some data

```
set.seed(704); n = 5000
income <- rnorm(n)#simulate independent income and bp data
bp <- rnorm(n)
gg <- ggplot(data.frame(income,bp), aes(income, bp)) +
  geom_point() + geom_smooth(method='lm', formula= y~x, color="darkmagenta", size=2) +
  labs(title = "No association of bp and income in population",
       subtitle = "Blue line is linear regression line") + theme_bw()
```



Plot the data

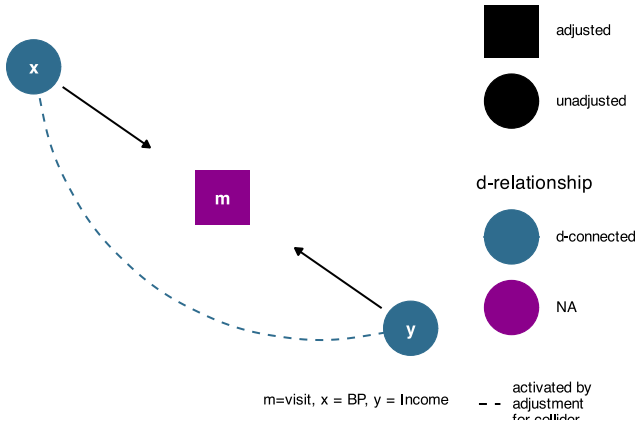


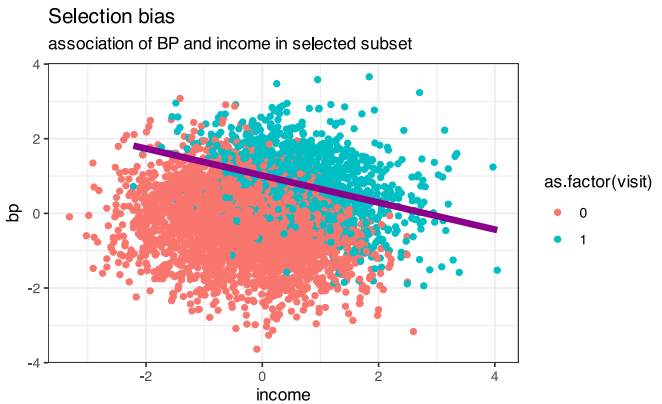
Understanding Selection Bias

What happens if "condition" on the common effect, i.e analyzing only visit=1

Let's simulate some data

```
library(tidyverse)
logitVisit <- -2 + 2*income + 2*bp # simulate visit = f(income, bp)
pVisit <- 1/(1+exp(-logitVisit)); visit <- rbinom(n, 1, pVisit);
dPop <- data.table::data.table(income, bp, visit)
# sample of those with a visit
dSample <- dPop[visit == 1]
```





In this conditioned sample there is now an association between BP and income

Understanding selection bias

Standard (Naive) approach

```
summary (lm(bp~income,
            data=dSample))
```

```
##
## Call:
## lm(formula = bp ~ income, data = dSample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4348 -0.5427 -0.0008  0.5079  3.3096
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.01237    0.02760   36.68   <2e-16 ***
## income      -0.36123    0.02447  -14.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.808 on 1418 degrees of freedom
## Multiple R-squared:  0.1333,    Adjusted R-squared:  0.1326
## F-statistic:   218 on 1 and 1418 DF,  p-value: < 2.2e-16
```

Remember p-values will not pick the causally correct model!

Understanding selection bias

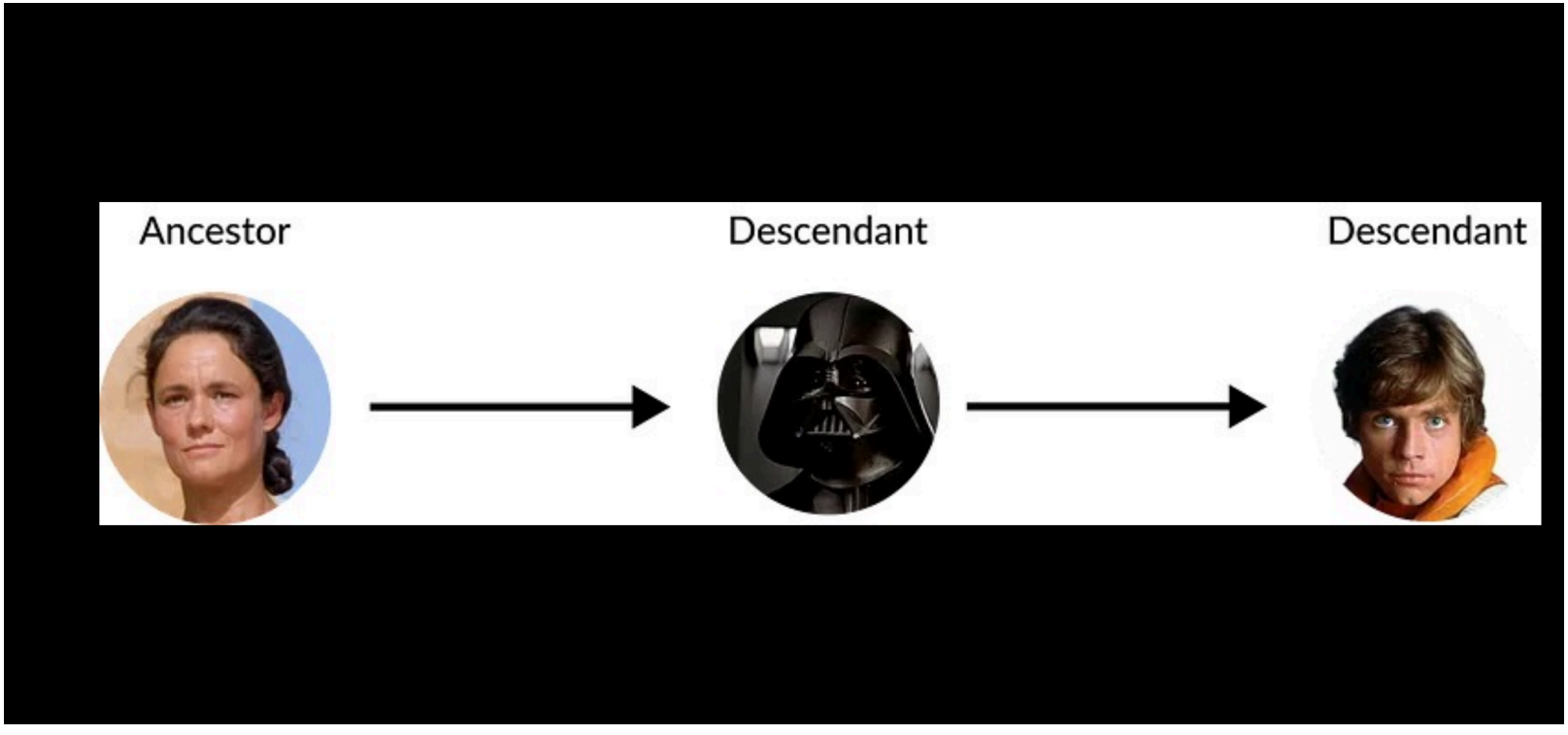
Model using all the data, which is not always available!

```
summary (lm(bp~income,
            data=dPop))
```

```
##
## Call:
## lm(formula = bp ~ income, data = dPop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6423 -0.6841 -0.0077  0.6759  3.6585
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.003972   0.014378   0.276    0.782
## income      -0.002509   0.014426  -0.174    0.862
##
## Residual standard error: 1.017 on 4998 degrees of freedom
## Multiple R-squared:  6.053e-06,    Adjusted R-squared:  -0.000194
## F-statistic: 0.03025 on 1 and 4998 DF,  p-value: 0.8619
```

Remember p-values will not pick the causally correct model!

Star DAGs



medium.com

Selection bias in RCTs

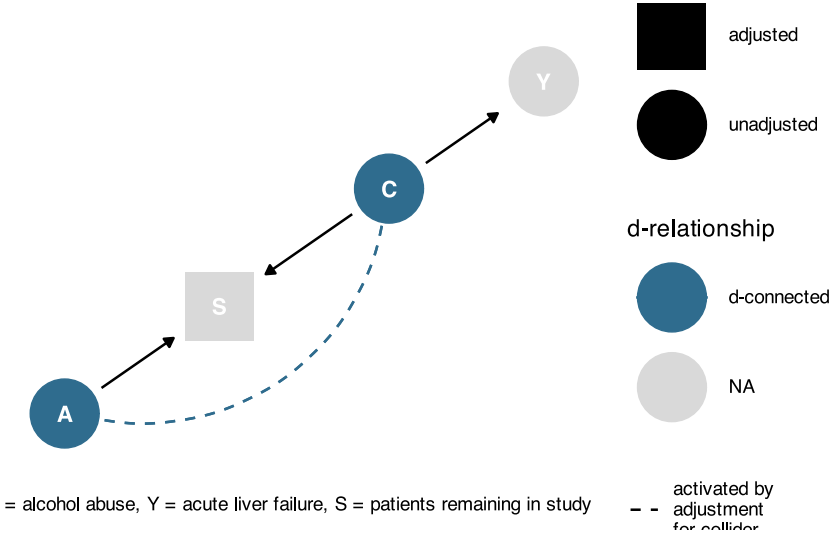
1. Can occur on entry if no blinding

2. Bigger issue is lost to follow-up

- Consider Rx (A) randomized but if (A=1); ↑ dropouts due to adverse drug effects (ADE). - Alcohol abuse (C=1) also more likely to drop out of the study and more likely to experience acute liver failure (Y=1).
- At baseline no association between A & C due to randomization

Selection bias in RCTs

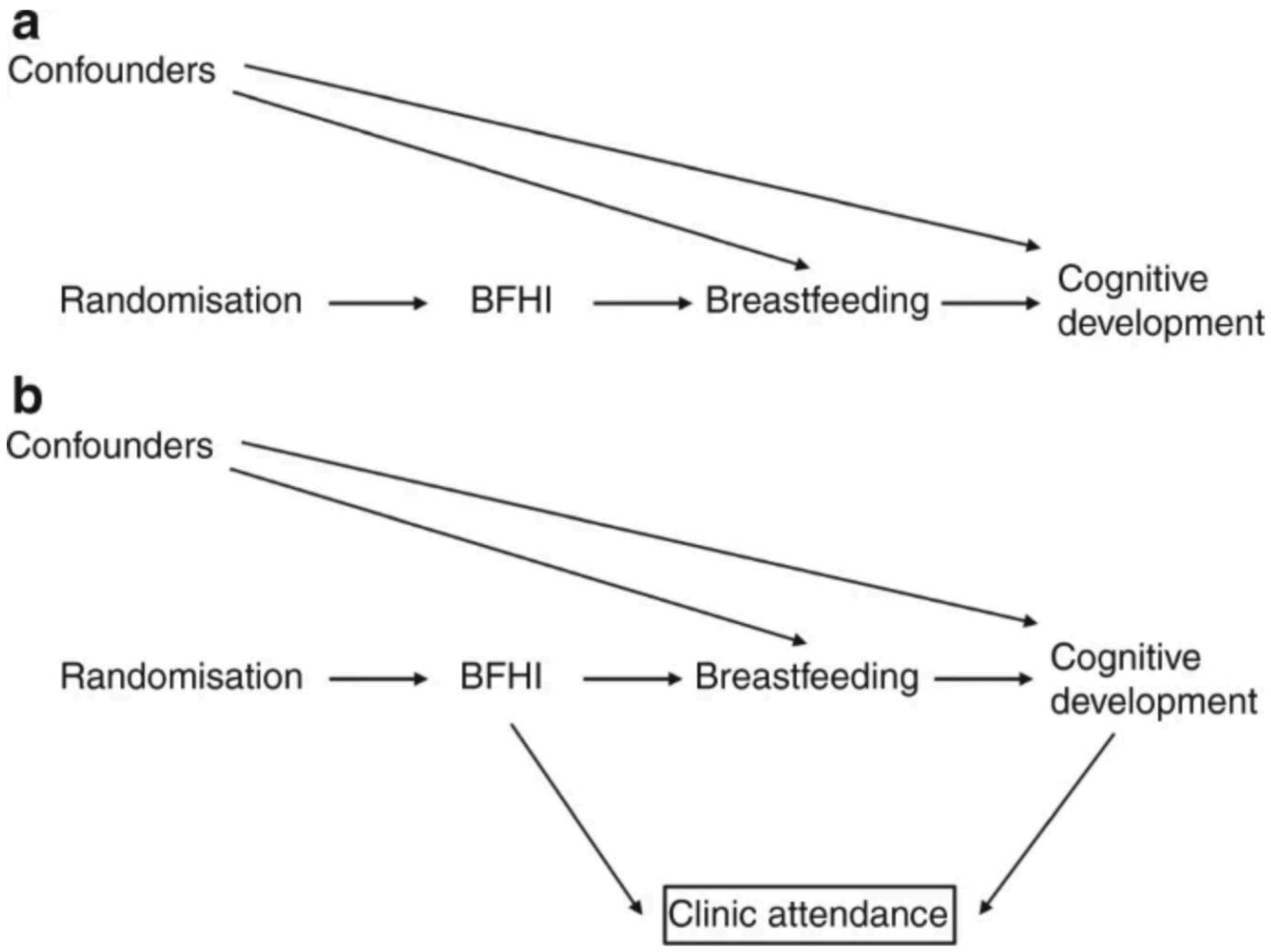
- At baseline no association between A & C due to randomization



- However, conditioning on those who do not drop out **creates a spurious backdoor negative correlation between A and C**

- Will make patients (A=1) to be protective against acute liver failure when no causal association exists.

Selection bias in RCTs



Directed acyclic graphs: a tool for causal studies in paediatrics

Summary collider vs. confounder

	Confounder	Collider
Main attribute	Common cause	Common Effect
Association	Contributes to the association between its effects	Does not contribute to the association od its effects
Type of path	Open path	Blocked path
Effect of conditioning	Blocks the path	Opens the path
Bias before conditioning?	Yes, confounding	No
Bias after conditioning?	No	Yes, Collider stratification Bias

Ever wonder about risk factor paradoxes?

Rheumatic diseases

Risk factor	Associations in the general population	Associations in the rheumatic disease (index) population
OA		
Bone mineral density	↑ Risk of incident OA	↓ Risk of OA progression[9]
Obesity	↑ Risk of incident OA	↔ Risk of OA progression[9]
Low vitamin C levels	↑ Risk of incident OA	↓ Risk of OA progression[9]
Female sex	↑ Risk of incident OA	↔ Risk of OA progression[9]
RA		
Smoking	↑ Risk of incident RA	↓ or ↔ Risk of RA progression[14–16]
	↑ Risk of incident CVD	↔ Risk of CVD among patients with RA[17–18]
Obesity	↑ Risk of mortality	↓ Mortality among patients with RA[20]
PsA		
Smoking	↑ Risk of psoriasis	↓ Risk of psoriatic arthritis among patients with psoriasis[4]
HLA-Cw*0602	↑ Risk of psoriasis	↓ Risk of psoriatic arthritis among patients with psoriasis[26,27]

Abbreviations: CVD, cardiovascular disease; OA, osteoarthritis; PsA, psoriatic arthritis; RA, rheumatoid arthritis.

Cardiac diseases

Risk factor paradox	Associations in the general population	Associations in the index population
Smoking paradox	↑ Risk of incident CAD	↓ Risk of hospital mortality in patients with CAD[28]
Obesity paradox	↑ Risk of incident CAD	↓ Risk of cardiovascular-specific mortality in patients with CAD[29,30]
Aspirin paradox	↑ Risk of incident COPD	↓ Mortality in patients with COPD[65]
Thrombophilia paradox	↑ Risk of incident CHD	↓ Risk of recurrent CHD events in patients with CHD[66]
PFO paradox	↑ Risk of incident VTE	↔ Risk of recurrent VTE in patients with incident VTE[35]
Low birth-weight paradox	↑ Risk of incident stroke ↑ Risk of low-birth weight baby	↔ Risk of recurrent stroke in patients with incident stroke[31,32] ↓ Mortality in low-birth weight babies
Apolipoprotein E4 allele	↑ Risk of incident Alzheimer disease	↓ Risk of Alzheimer disease progression[33,34]

Abbreviations: CAD, coronary artery disease; CHD, coronary heart disease; COPD, chronic obstructive pulmonary disease; PFO, patent foramen ovale; VTE, venous thrombotic embolism.

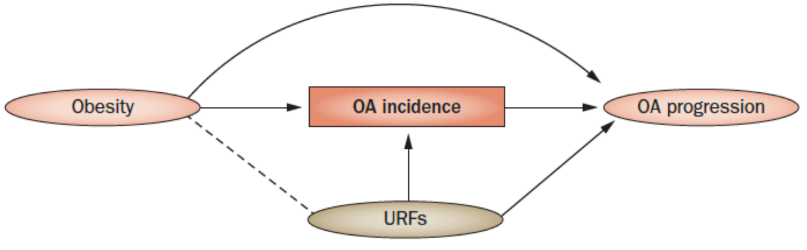
Choi, H. K. et al. Nat. Rev. Rheumatol. 10, 403–412 (2014); published online 1 April 2014; doi:10.1038/nrrheum.2014.36

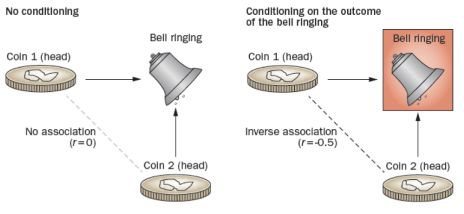
Well established risk factors in general population reverse their impact in these selected index populations???

What's going on?

Editors like the word “paradox” and its mention increases likelihood of publication - novel, controversial findings, easy to invent hypothetical explanations

- But is this a causal or non-biological explanation?





Remember stratifying on a collider $\rightarrow$ spurious negative association among those risk factors in indexed (stratified) populations as the most likely explanation with an index event

When you see the word paradox, think first about Index event (collider stratification) bias

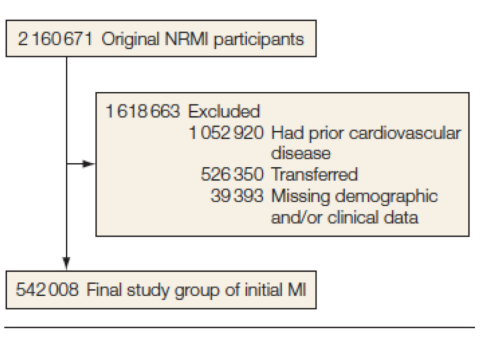
An egregious example (with a dose response!)

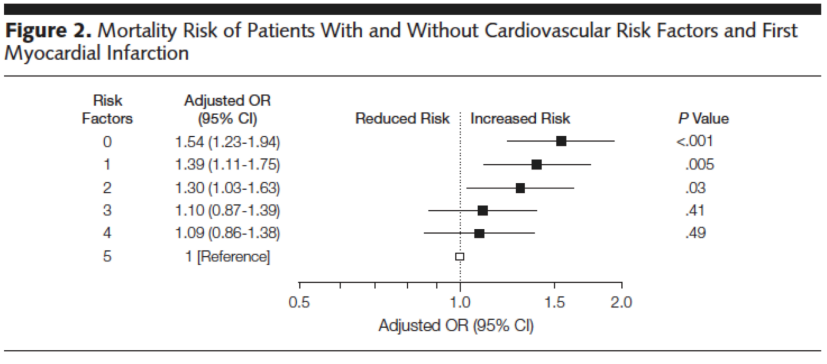
The following was published in JAMA

Number of Coronary Heart Disease Risk Factors and Mortality in Patients With First Myocardial Infarction

Conclusion Among patients with incident acute myocardial infarction without prior cardiovascular disease, in-hospital mortality was inversely related to the number of coronary heart disease risk factors.

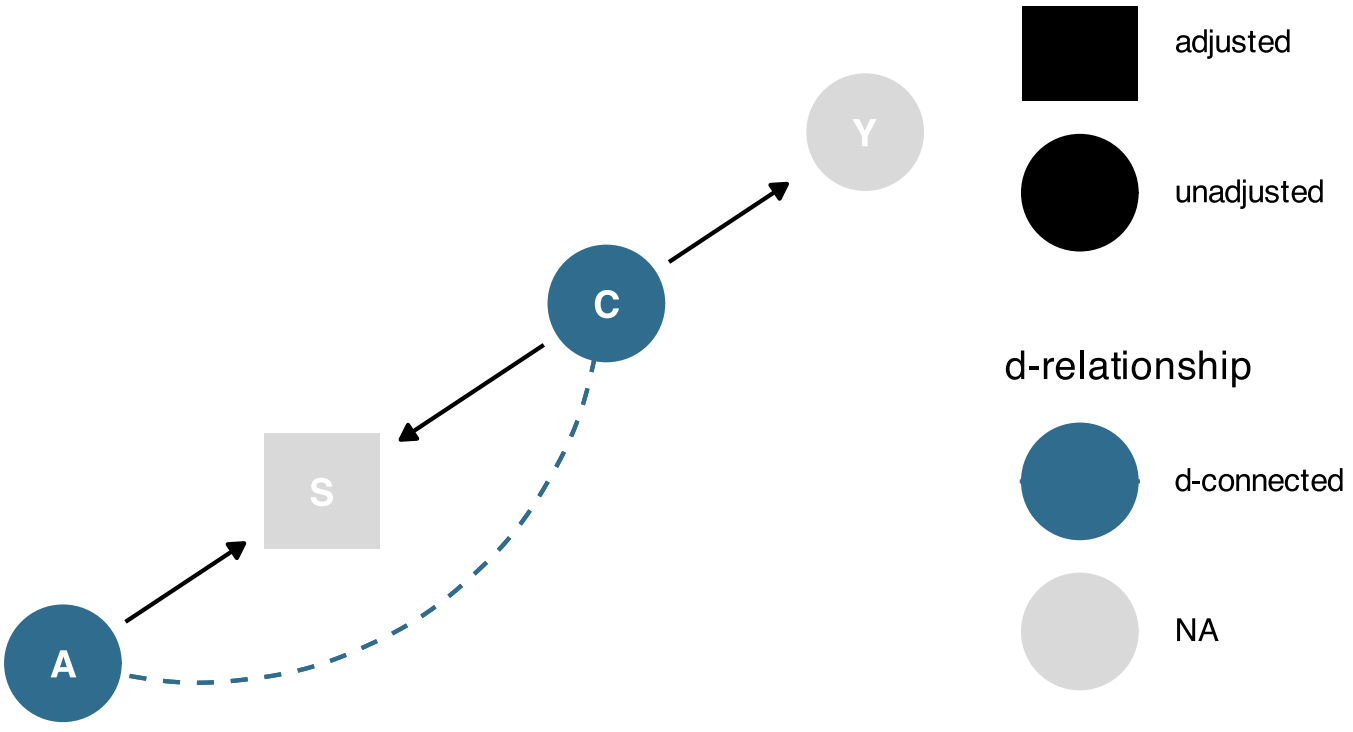
JAMA. 2011;306(19):2120-2127 www.jama.com





Should we encourage post MI patients to increase their smoking, weight, cholesterol, BP and diabetes?... And ideally do all of the above simultaneously **REALLY!?**

DAGs with R



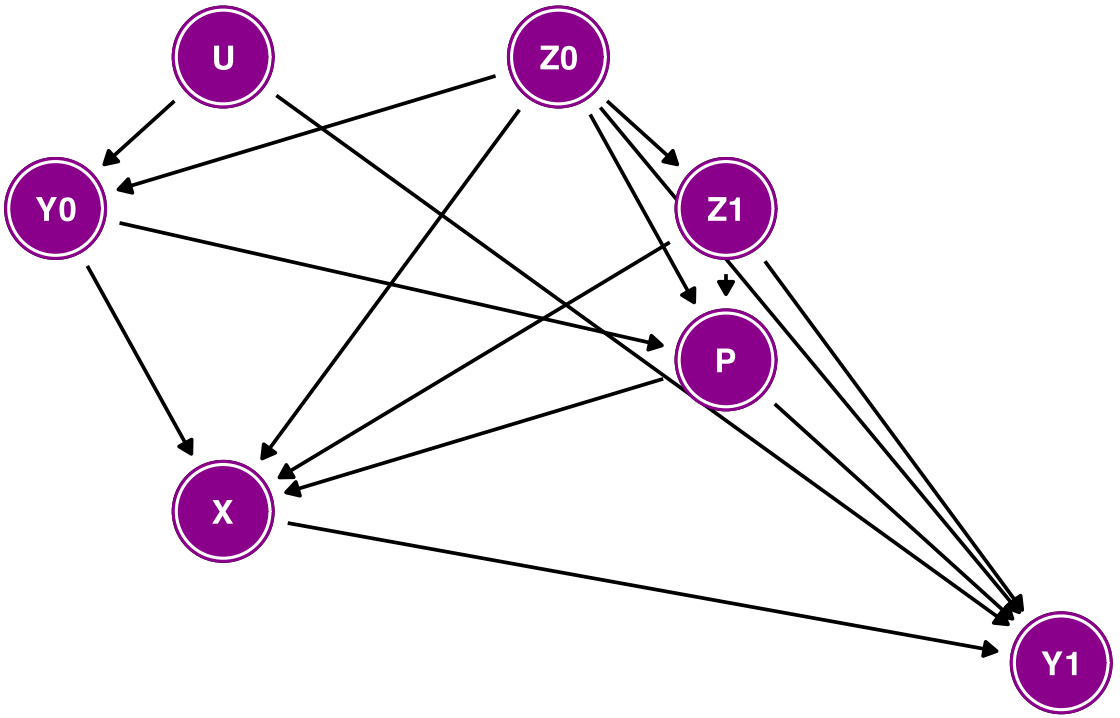
A = Rx, C = alcohol abuse, Y = acute liver failure,
S = patients remaining in study

– – activated by adjustment for collider

DAGs with R

```r
library(ggdag) # help(package="ggdag)
l3dag1<- dagify(S ~ A + C, Y ~ C) %>% #dagify() creates dagitty DAGs using a R-like syntax.
  tidy_dagitty() %>%
  node_dconnected("A", "C", controlling_for = "S") %>%
  ggplot(aes(
    x = x,
    y = y,
    xend = xend,
    yend = yend,
    shape = adjusted,
    col = d_relationship )) +
  geom_dag_edges(aes(end_cap = ggraph::circle(10, "mm"))) +
  geom_dag_collider_edges() +
  geom_dag_point() +
  geom_dag_text(col = "white") +
  theme_dag() +
  scale_adjusted() +
  expand_plot(expand_y = expansion(c(0.2, 0.2))) +
  scale_color_viridis_d(
    name = "d-relationship",
    na.value = "grey85",
    begin = .35 ) +
  labs(caption = "A = Rx, C = alcohol abuse, Y = acute liver failure,
       S = patients remaining in study")
l3dag1
```

A more complicated DAG



How to determine the causal effect of X on Y1?

A more complicated DAG

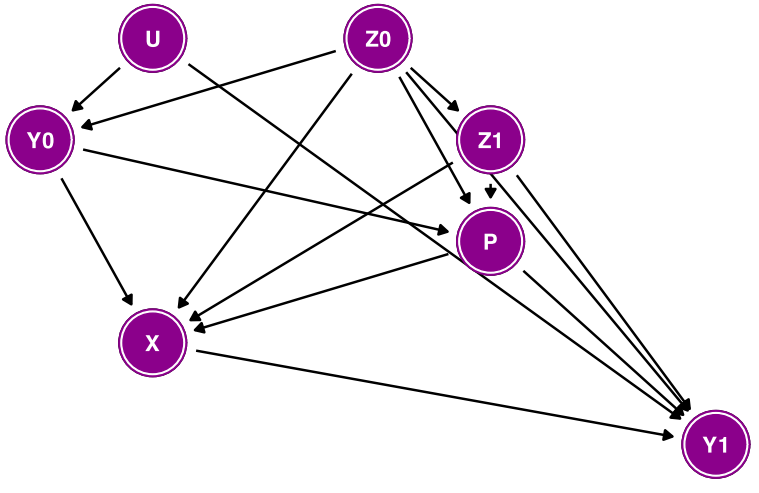
```
dag <- ggdag::dagify(Y1 ~ X + Z1 + Z0 + U + P,
                     Y0 ~ Z0 + U,
                     X ~ Y0 + Z1 + Z0 + P,
                     Z1 ~ Z0,
                     P ~ Y0 + Z1 + Z0,
                     exposure = "X",
                     outcome = "Y1")
dag_plot <- dag %>%
  ggdag::tidy_dagitty(layout = "auto", seed = 12345) %>%
  arrange(name) %>%
  ggplot(aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_dag_point() +
  geom_dag_edges() +
  geom_dag_text(parse = TRUE, label = c("P", "U", "X",
                        expression(Y[0]), expression(Y[1]), expression(Z[0]), expression(Z[1])
  theme_dag() +
  geom_dag_node(color="darkmagenta") + geom_dag_text(color="white")

dag_plot
```

How to determine the causal effect of X on Y1?

R can help

Questions arising from this DAG.

1. How many paths are there from X to Y1?
2. How many of those paths are spurious (backdoor) paths?
3. How many of those backdoor paths are open?
4. What is the minimal set of variables to block these spurious pathways?



Questions theoretically answerable by careful attention to DAG but easier with `dagitty` built-in functions

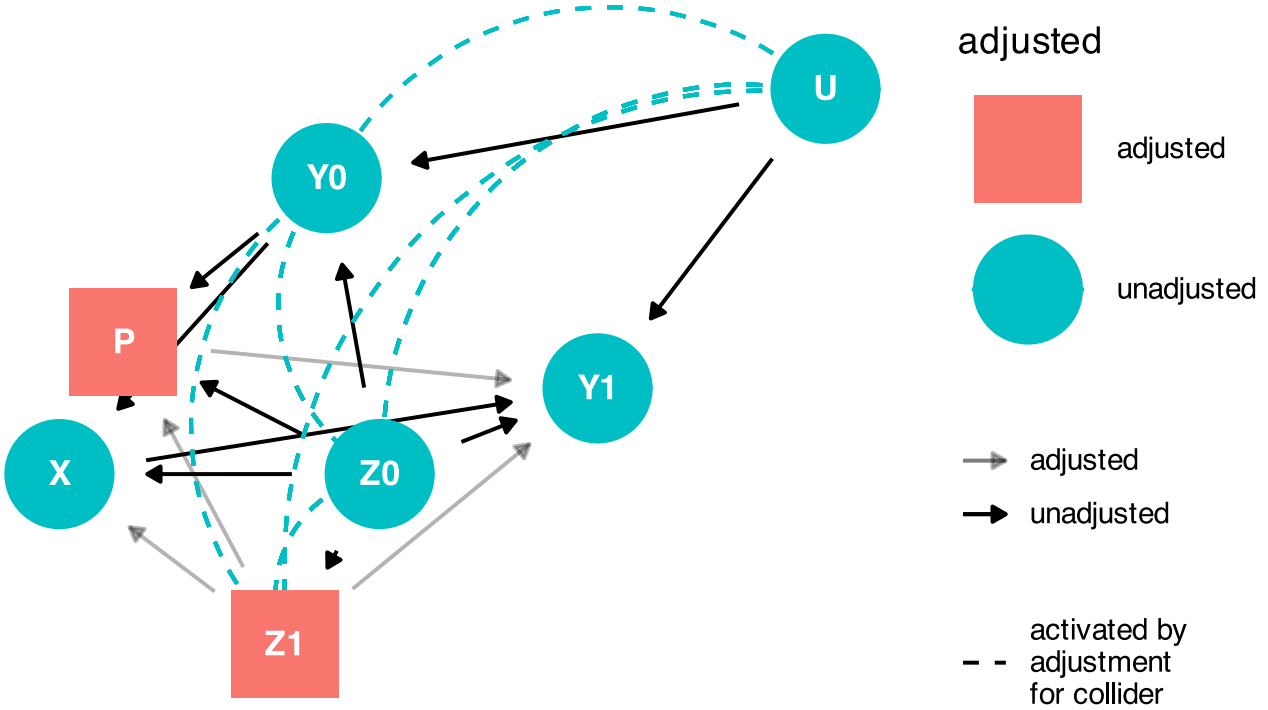
R can help ...

```
g <- dagitty::paths(dag, "X", "Y1")
a <- paste0("There are ", length(g$paths),
            " pathways from X to Y1 and all are backdoor except for 1")
b <- paste0("Of these backdoor pathways ",
            sum(g$open=="TRUE"), " are open")
c <- paste0("The minimum adjustment sets are ",
            adjustmentSets(dag, "X", "Y1", type = "minimal"))

print(c(a,b,c))
```

```
## [1] "There are 43 pathways from X to Y1 and all are backdoor except for 1"
## [2] "Of these backdoor pathways 25 are open"
## [3] "The minimum adjustment sets are c(\"P\", \"U\", \"Z0\", \"Z1\")"
## [4] "The minimum adjustment sets are c(\"P\", \"Y0\", \"Z0\", \"Z1\")"
```

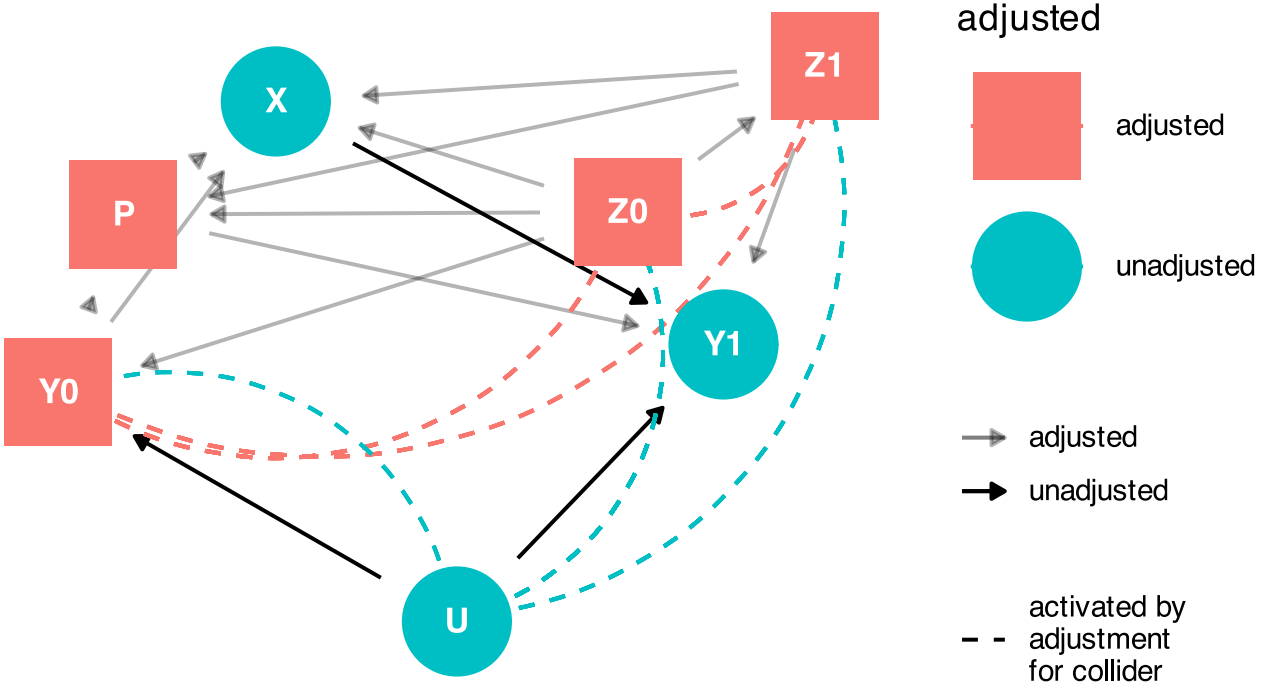
"Incomplete" Adjustment

```
ggdag_adjust(dag, var = c("Z1", "P"))+ theme_dag()
```



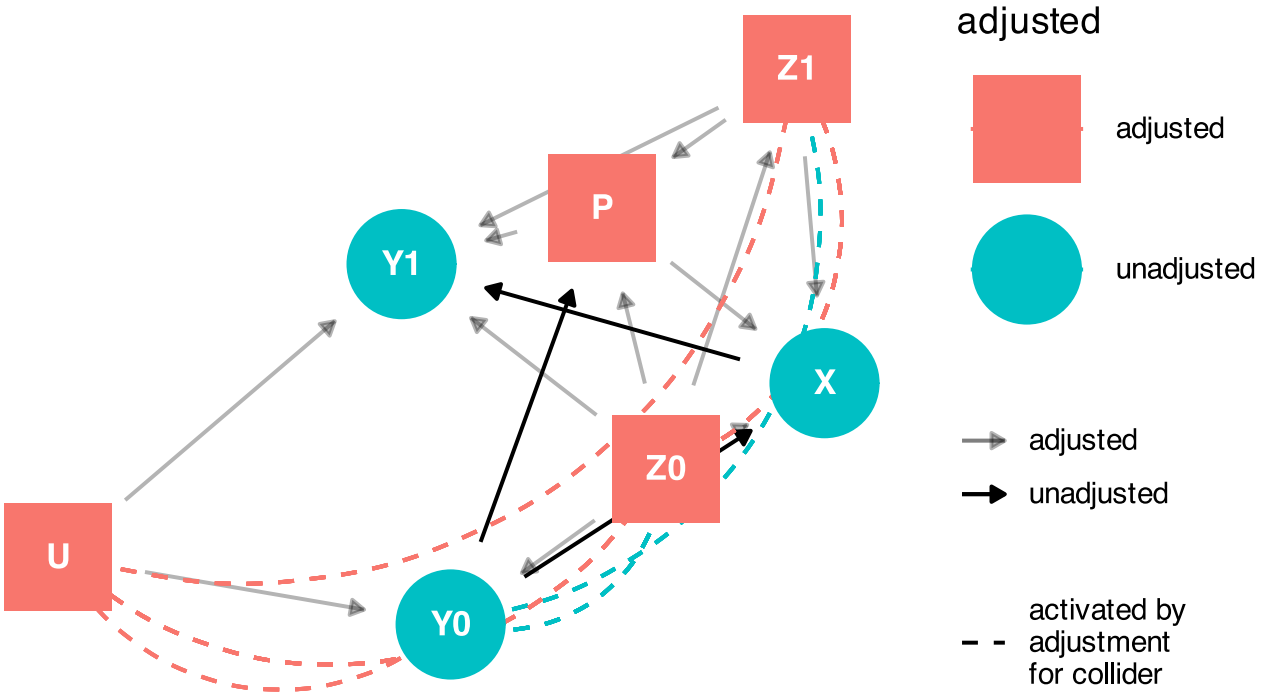
"Complete" Adjustment

```
ggdag_adjust(dag, var = c("Z1", "Z0", "P", "Y0"))+ theme_dag()
```



"Complete" Adjustment (but 'U' was not labeled as 'latent')

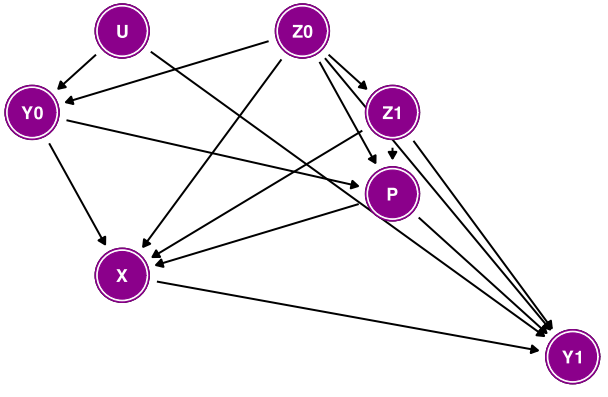
```
ggdag_adjust(dag, var = c("Z1", "Z0", "P", "U"))+ theme_dag()
```



Labelling the 'U' as latent

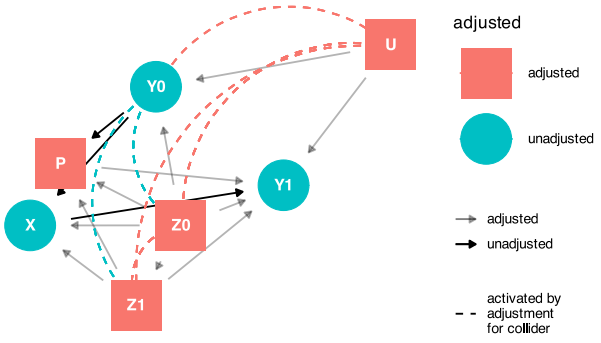
```
dag0 <- ggdag::dagify(Y1 ~ X + Z1 + Z0 + U
                      Y0 ~ Z0 + U,
                      X ~ Y0 + Z1 + Z0 + P,
                      Z1 ~ Z0,
                      P ~ Y0 + Z1 + Z0,
                      exposure = "X",
                      outcome = "Y1",
                      latent = "U")
dag_plot0 <- dag0 %>%
  ggdag::tidy_dagitty(layout = "auto",
                      seed = 12345) %>%
  arrange(name) %>%
  ggplot(aes(x = x, y = y,
             xend = xend, yend = yend)) +
  geom_dag_point() +
  geom_dag_edges() +
  geom_dag_text(parse = TRUE,
                label = c("P", "U", "X",
                expression(Y[0]), expressio
  theme_dag() +
  geom_dag_node(color="darkmagenta") +
  geom_dag_text(color="white")
```

Does not change the DAG



But the revised DAG has a revised "Complete" Adjustment set

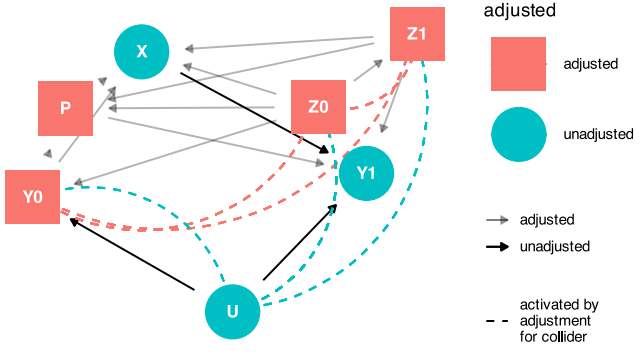
```
ggdag_adjust(dag0,
             var = c("Z1", "Z0", "P", "U"))
  theme_dag()
```



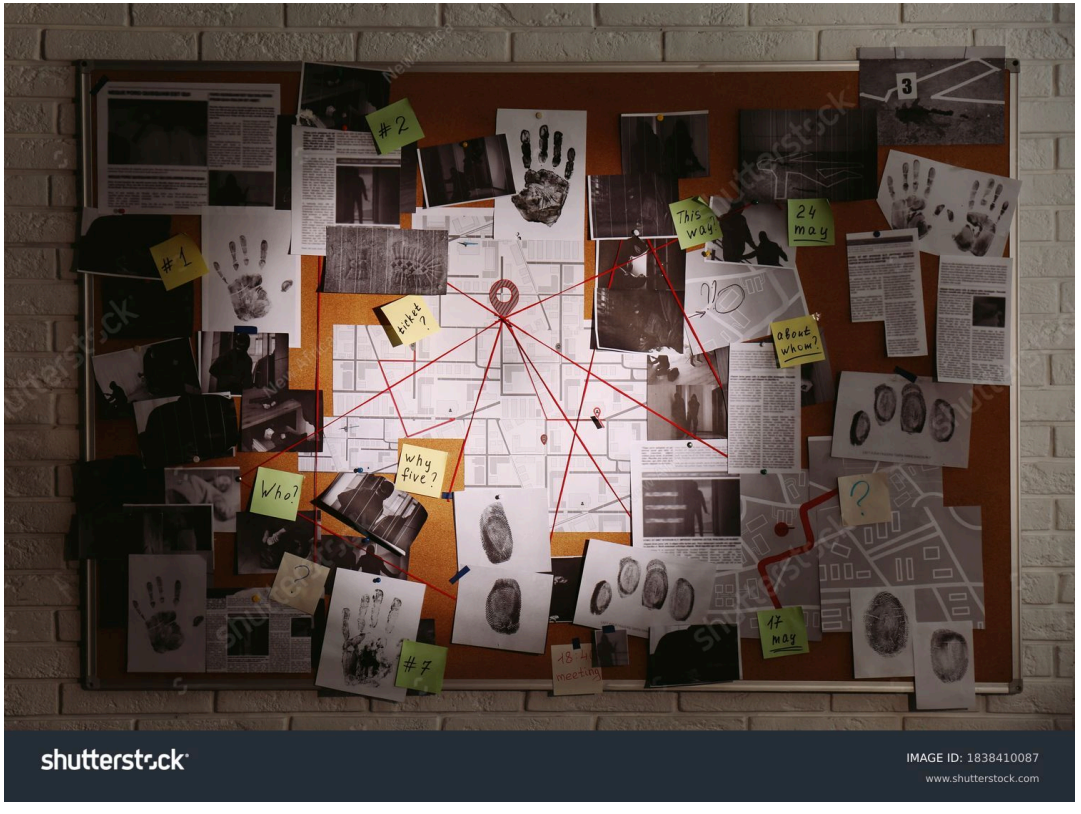

```
adjustmentSets(dag0, "X", "Y1", type = "min
```

```
## { P, Y0, Z0, Z1 }
```

```
ggdag_adjust(dag0,
             var = c("Z1", "Z0", "P", "Y0")
  theme_dag()
```



Ready to solve the case?



Conclusion: DAGs can be super useful on the road to causal inference

References

- Williams, T.C., Bach, C.C., Matthiesen, N.B. et al. Directed acyclic graphs: a tool for causal studies in paediatrics. Pediatr Res 84, 487–493 (2018). https://doi.org/10.1038/s41390-018-0071-3
- Haber NA, Wood ME, Wieten S, Breskin A. DAG With Omitted Objects Displayed (DAGWOOD): a framework for revealing causal assumptions in DAGs. Ann Epidemiol. 2022 Apr;68:64-71. https://doi.org/10.1016/j.annepidem.2022.01.001
- Piccininni, M., Konigorski, S., Rohmann, J.L. et al. Directed acyclic graphs and causal thinking in clinical risk prediction modeling. BMC Med Res Methodol 20, 179 (2020). https://doi.org/10.1186/s12874-020-01058-z
- Pearl J. An introduction to causal inference. Int J Biostat. 2010 Feb 26;6(2):Article 7. doi: 10.2202/1557-4679.1203. PMID: 20305706; PMCID: PMC2836213.
- Causal Inference with R (https://www.r-causal.org/chapters/05-dags)
- Causal Diagrams website: https://causaldiagrams.org/

Extra Resources:

- https://dagitty.net/
- dagshub.com
- medium.com
- Causal Inference Ch-5
- Pearl, Judea. Causal Inference in Statistics : A Primer, John Wiley & Sons, Incorporated, 2016. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/mcgill/detail.action?docID=7104473.

QUESTIONS?

COMMENTS?

RECOMMENDATIONS?

Some extra tips on DAGs

DAGs with R

Key functions are `dagify()` in `ggdag` package
Consider the model: `smoking` causes `cancer` and `addictive` behaviour cases both `coffee` drinking and `smoking` but `coffee` does not cause `cancer`

Create the dagitty object with `dagify`

```
coffee_cancer_dag <- ggdag::dagify(
  cancer ~ smoking,
  smoking ~ addictive,
  coffee ~ addictive,
  exposure = "coffee",
  outcome = "cancer",
  labels = c("coffee" = "Coffee",
             "cancer" = "Lung cancer",
    "smoking" = "Smoking",
    "addictive" = "Addictive Behavior"
  ))
class(coffee_cancer_dag )
```

```
## [1] "dagitty"
```

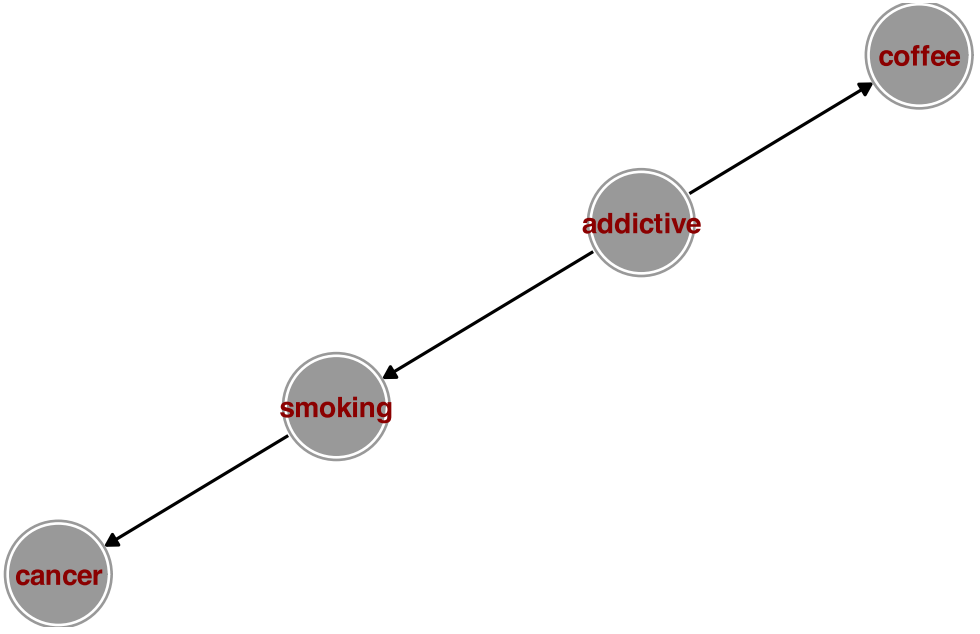
```
coffee_cancer_dag
```

```
## dag {
## addictive
## cancer [outcome]
## coffee [exposure]
## smoking
## addictive -> coffee
## addictive -> smoking
## smoking -> cancer
## }
```

Plot the daggity object

Use `ggdag()` in ggdag package

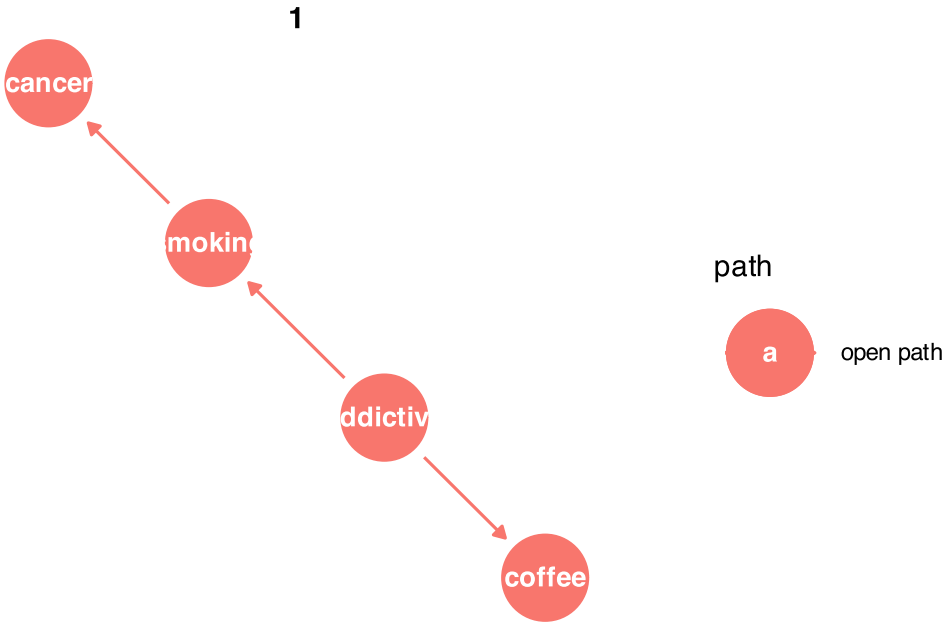
```
ggdag(coffee_cancer_dag) +
  #geom_dag_edges(aes(end_cap = ggraph::circle(14, "mm"))) +
  ggdag::geom_dag_node(color="grey60", size = 20) +
  geom_dag_point(color="grey60") +
  geom_dag_text(col = "darkred") +
  theme_dag()
```



Open pathways

Can be determined automatically with `ggdag_paths()`

```
coffee_cancer_dag %>%
  ggdag_paths() +
  #ggdag::geom_dag_node(color="grey60", size = 20) +
  #geom_dag_text(col = "darkred") +
  theme_dag()
```



Closing backdoor paths

- Randomization, regression, stratification, weighting, matching
- Identifying variable for adjusting with R

```
ggdag_adjustment_set(coffee_cancer_dag, use_labels = "label", text = FALSE)+ theme_dag()
```