Lösungen zu den Aufgaben

1. **Aufgabe**

   Eine Forscher:in aus Kalifornien entdeckt, dass Haiangriffe mit Eisverkauf korreliert sind: Haiangriffe treten gehäuft dann auf, wenn am Strand viel Eis verkauft wird. Dieser Zusammenhang ist zwar nicht perfekt, aber die Forscher:in findet in ihren Daten einen starken, sogar "signifikanten" Zusammenhang.

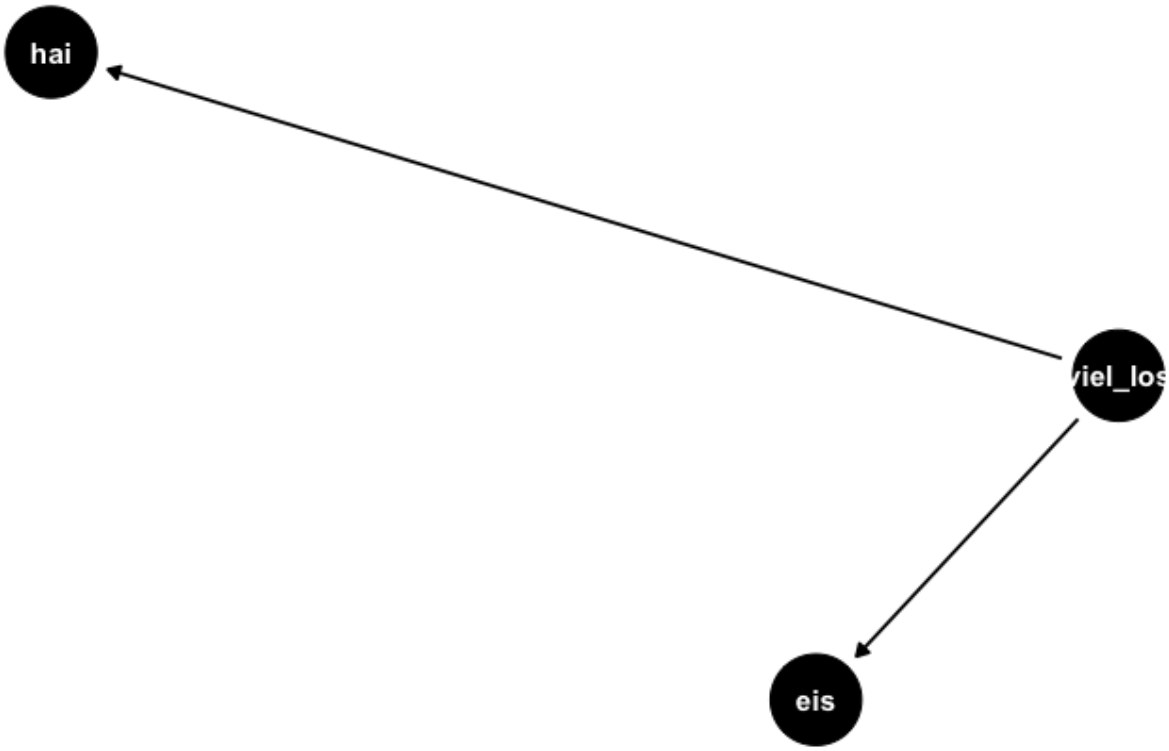
   Welche Schlüsse sind aus diesen Daten zu ziehen? Wählen Sie die Antwort, die am besten passt!

   a. Da *Eisverkauf* die UV und *Haiangriff* die AV ist, sind die Daten im Sinne eines Kausalschlusses "Eisverkauf führt (tendenziell) zu Haiangriffen" zu interpretieren. Natürlich gilt dies nur für linearen Zusammenhänge, da Korrelationen nur linearen Zusammenhänge identifizieren können.
   b. Es ist kein Kausalschluss möglich; eine Drittvariable könnte den Zusammenhang der beobachteten Variablen konfundieren.
   c. Die Daten (soweit bekannt bzw. oben aufgeführt sind) machen deutlich, dass es einen Zusammenhang zwischen den beiden Variablen gibt; folglich ist die eine Variable Ursache und die andere Wirkung. Die Daten lassen aber keine Aussage zu, welche der beiden Variablen Ursache und welche Wirkung ist.
   d. Es ist davon auszugehen, dass *Haiangriff* die Ursache ist und *Eisverkauf* die Wirkung.
   e. Da es sich nur um Beobachtungsdaten, nicht um Experimentaldaten handelt, ist keine Aussage möglich.

   Lösung

   Es ist kein Kausalschluss möglich; eine Drittvariable könnte den Zusammenhang der beobachteten Variablen konfundieren.

   Diese Drittvariable könnte das Aufkommen der Besucher:innen am Strand sein (`viel_los`). Wenn viel los ist am Strand, steigt die Gefahr an Haiangriffen, einfach weil mehr Menschen im Wasser sind. Weiter gilt: Wenn viel los ist, wird viel Eis verkauft. Diese beiden Kausaleffekte lassen eine Scheinkorrelation zwischne `hai` und `eis` erscheinen: Scheinbar gibt es einen Kausaleffekt zwischen Eisverkauf und Haiangriffen. Dieser Zusammenhang ist aber eine Scheinkorrelation, kein Kausaleffekt. Das Diagramm zeigt diese Konfundierung.



a. Falsch
b. Wahr
c. Falsch
d. Falsch
e. Falsch

2. **Aufgabe**

Denken wir uns ein kausales System mit einer Ursache und einer Wirkung, etwa der Einfluss der Naturbelassenheit (N) eines Landkreises auf die Anzahl der Störche (S) dort (ein positiver Einfluss). Nehmen wir weiter an, die Naturbelassenenheit eines Landkreises hat einen (positiven) Einfluss auf die Anzahl Neugeborener (Babies, B).

Weitere kausale Einflüsse existieren in diesem kausalen System nicht (es handelt sich ja hier umn ein Gedankenexperiment, wir können frei bestimmen!).

Die Frage ist nun, ob wir erwarten müssen, dass Störche und Babies zusammenhängen in diesem System, dass es also dort, wo es viele Störche gibt auch viele Babies gibt. Das wäre deswegen beachtlich, weil wir in unserem System explizit keinen (kausalen) Zusammenhang zwischen diesen beiden Größen definiert haben.

Um die Sache etwas greifbarer zu machen, erstellen wir uns Daten, die zu diesem System passen. Sagen wir, wir haben 100 Landkreise, die in der Zahl der Störche und Babies und Naturbelassenheit variieren. Der Einfachheit halber seien alle Werte in z-Werten ausgedrückt. Gehen wir weiter (der Einfachheit halber) davon aus, alle Größen sind normalverteilt. Solche Werte kann man mit der R-Funktion `rnorm()` erzeugen.

Schließlich gehen wir noch davon aus, dass die Einflüsse linear sind und nicht perfekt. Der Zufall (zufälliger “Fehler”, e) soll also auch einen Einfluss auf die Größen haben.

```
N <- rnorm(100, mean = 0, sd = 1)  # 100 normalverteilte z-Werte
e1 <- rnorm(100)  # das gleiche wie oben: normalverteilte z-Werte
e2 <- rnorm(100)  # das gleiche wie oben: normalverteilte z-Werte
S <- N + e1  # S wird determiniert durch N und e
B <- N + e2  # B wird determiniert druch N und e
```

Testen wir unsere simulierten Daten mit einer einfachen Regression, der Frage, ob die Anzahl der Störche (S) von der Natürlichkeit (N) abhängt:

```
lm1 <- lm(S ~ N)
summary(lm1)
```

```
##
## Call:
## lm(formula = S ~ N)
##
## Residuals:
##      Min       1Q   Median       3Q
## -2.35998 -0.81570  0.03617  0.61823
##      Max
##  2.62146
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)  0.01993    0.11015   0.181
## N            0.90983    0.12931   7.036
##             Pr(>|t|)
## (Intercept)    0.857
## N           2.71e-10 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
##   0.1 ' ' 1
##
## Residual standard error: 1.101 on 98 degrees of freedom
## Multiple R-squared:  0.3356, Adjusted R-squared:  0.3288
## F-statistic: 49.51 on 1 and 98 DF,  p-value: 2.707e-10
```

Unser Modell `lm1` bringt unsere Annahmen deutlich zum Vorschein.

a. Bestimmen Sie den Zusammenhang (β oder ρ) zwischen Störchen und Babies!
b. Erklären Sie den Befund!

Lösung

a. Es findet sich ein nicht-kausaler, also ein *Scheinzusammenhang* zwischen Störchen und Babies:

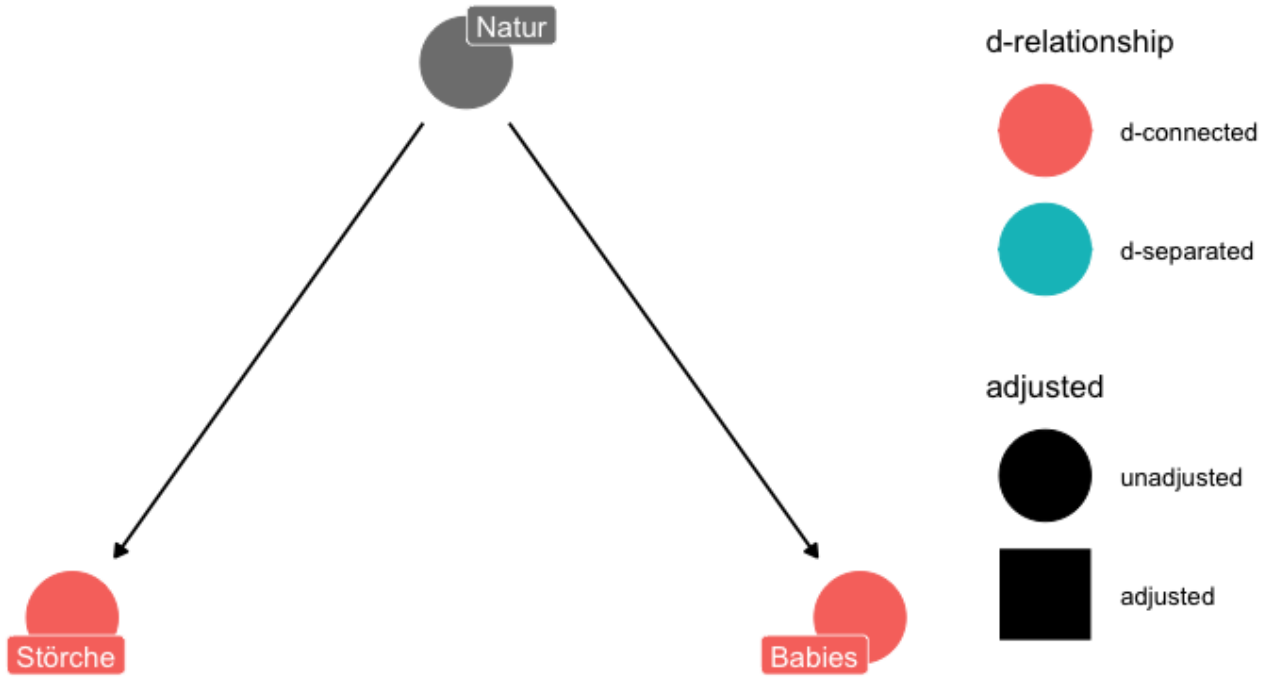
```
lm(B ~ S)
```

```
##
## Call:
## lm(formula = B ~ S)
##
## Coefficients:
## (Intercept)            S
##      0.1129       0.3430
```

```
cor(S, B)
```

```
## [1] 0.3518875
```

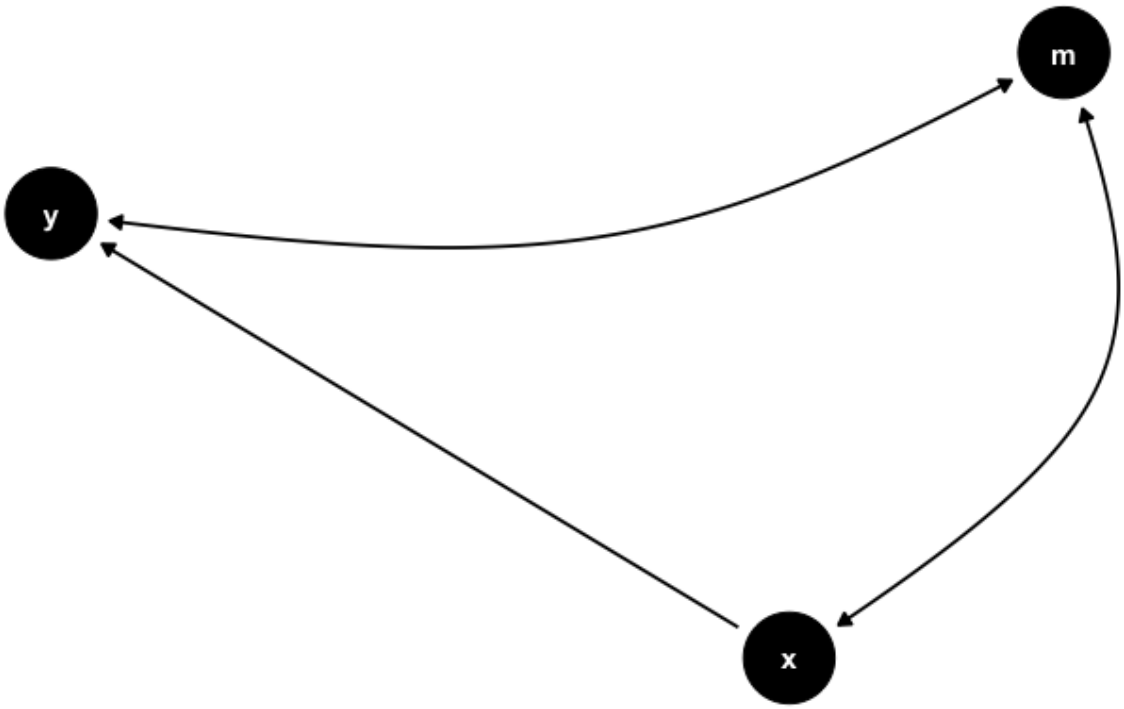
b. Haben zwei Variablen eine gemeinsame Ursache, so sind sie durch eine Scheinkorrelation verbunden.



`dconnected` bedeutet, dass zwei Variablen *verbunden* (connected) sind, sie also voneinander (statistisch) abhängig (assoziiert) sind, z.B. korreliert. Das `d` steht für `directed` also über gerichtete Kanten, die Kausalpfeile, verbunden. Dabei ist zu beachten, dass die Assoziation in beide Richtungen des Kausalpfeils “fließen” kann; auch gegen “den Strom” (also von der Pfeilstütze anfangend rückwärts).

3. **Aufgabe**

DAG g



Gegeben sei der DAG `g` (s.o.). Dabei ist zu beachten, dass die gebogene Kurve (keine Gerade) mit zwei Pfeilspitzen *keinen* Kausaleffekt beschreibt, sondern eine *Assoziation.* Die dahinterstehende kausale Struktur ist eine Konfundierung. Daher ist der “Doppelpfeil” als Abkürzung für eine Konfundierung zu verstehen.

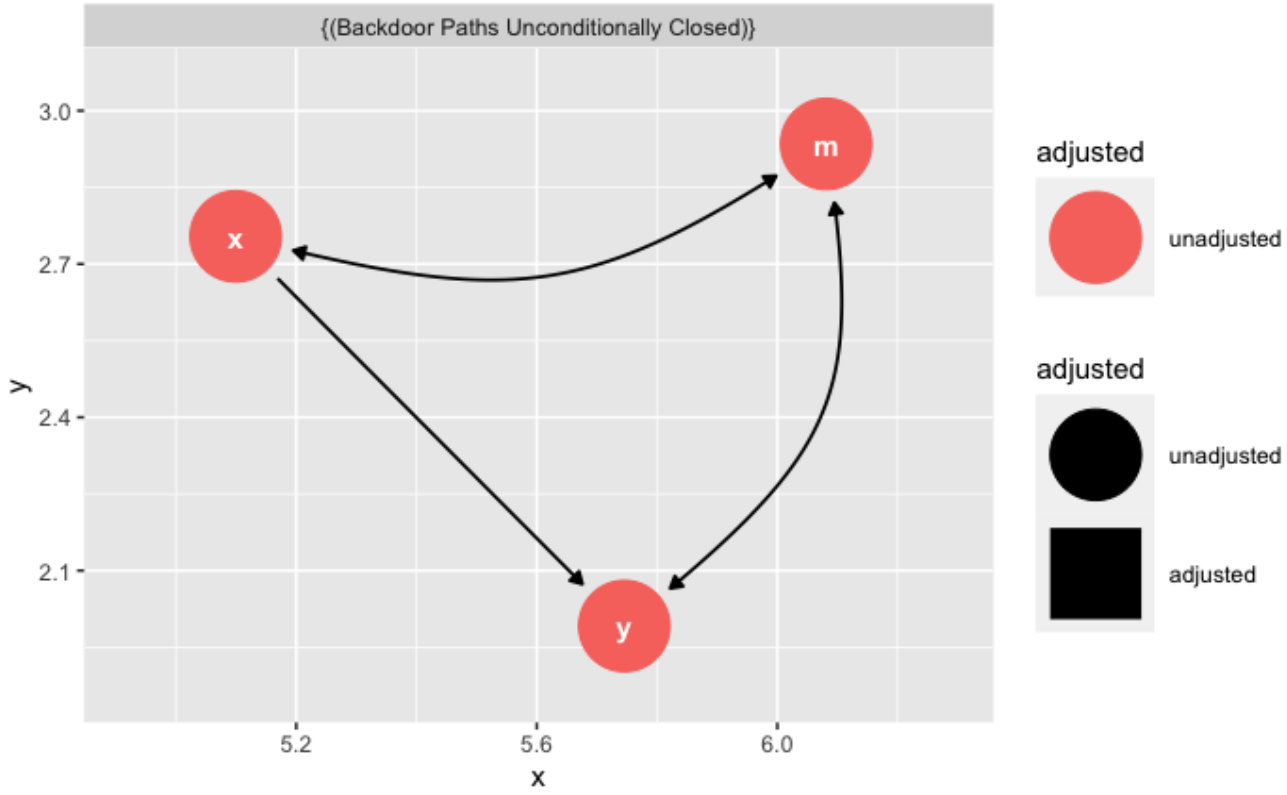
Welche Variable/n sind zu kontrollieren, um den kausalen Effekt von `x` auf `y` zu identifizieren?

a. x
b. y
c. keine, bereits identifiziert
d. m
e. keine, nicht identifizierbar

Lösung

Keine. Der kausale Effekt von `x` auf `y` ist bereits identifiziert. Identifiziert bedeutet, dasss die statistische Assoziation zwischen den beiden Variablen komplett kausal ist. Es gibt keine Hintertürpfade.

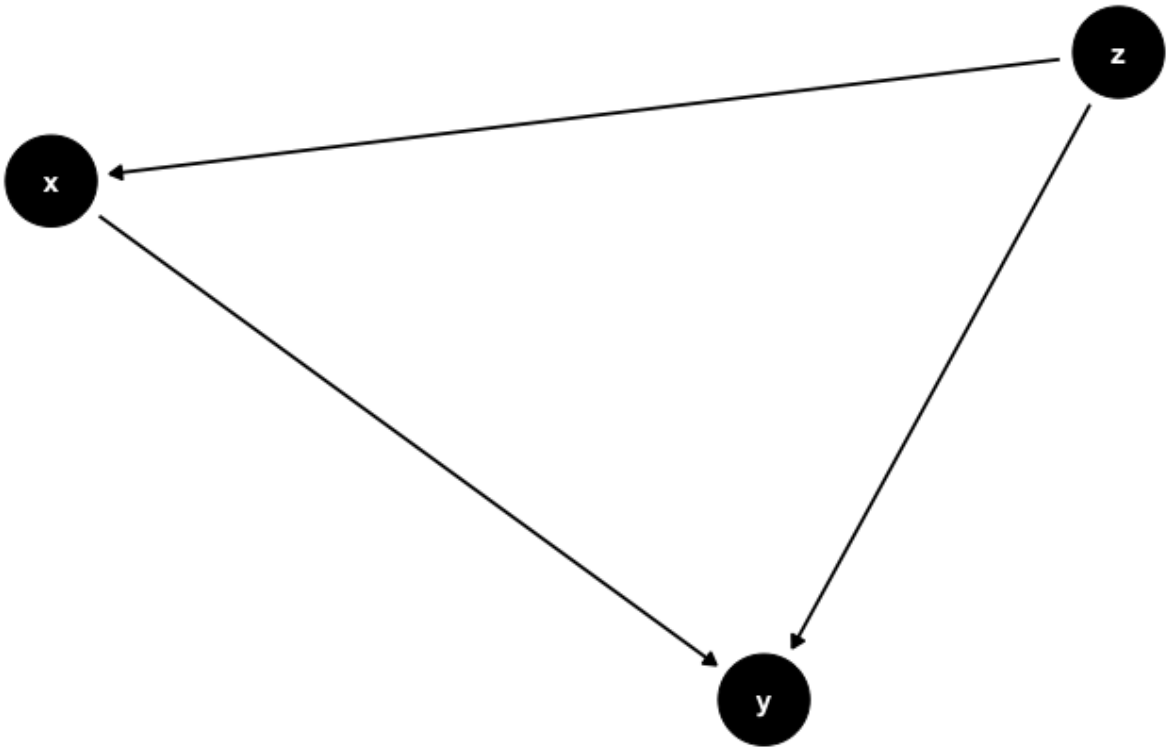
```
##  {}
```



a. Falsch
b. Falsch
c. Wahr
d. Falsch
e. Falsch

4. **Aufgabe**

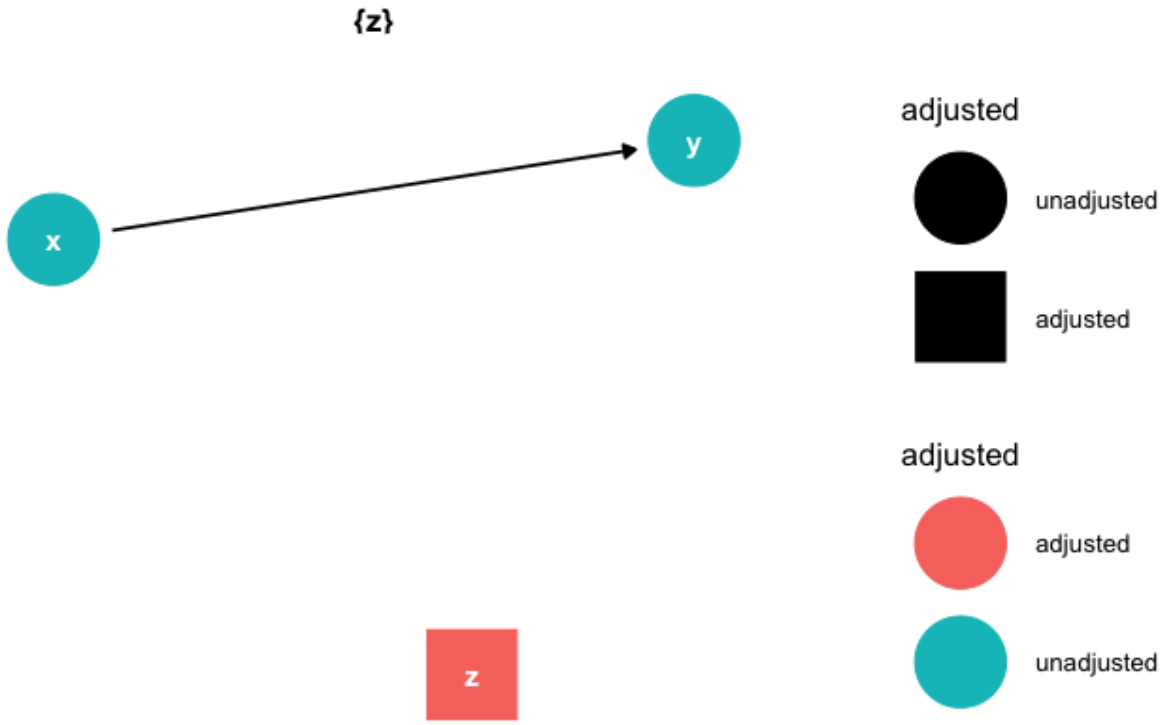
Gegeben sei der DAG `g` (s.u.). Welche Variable/n sind zu kontrollieren, um den kausalen Effekt von `x` auf `y` zu identifizieren?



a. keine, bereits identifiziert
b. x
c. y
d. keine, nicht identifizierbar
e. z

Lösung

Durch Kontrolle von `z` wird der kausale Effekt von `x` auf `y` identifiziert.

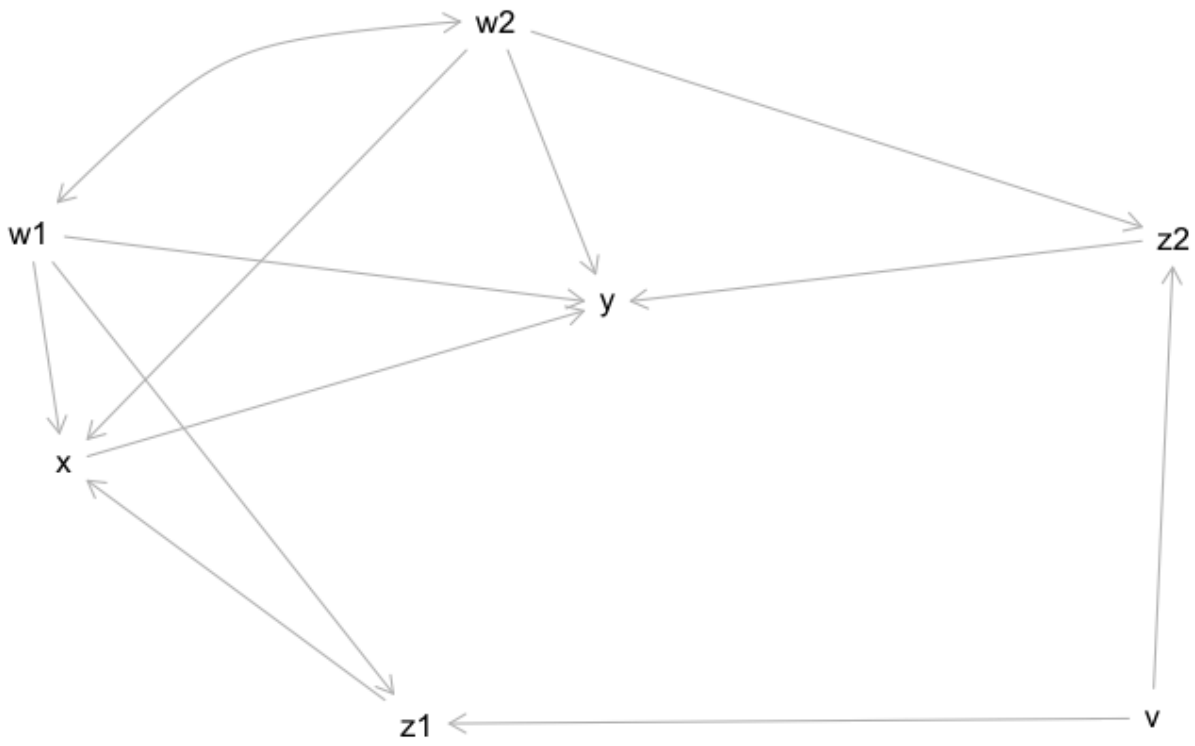


Anmerkung: *Kein* Pfeil bedeutet, dass kein kausaler Pfad geöffnet ist.

a. Falsch
b. Falsch
c. Falsch
d. Falsch
e. Wahr

5. **Aufgabe**

Gegeben sei der DAG `g` (s.u.). Was ist die minimale Menge an Variablen, die man kontrollieren muss, um den kausalen Effekt von `x` auf `y` zu identifizieren?



Hinweise:

- Gebogene Kurven mit doppelter Pfeilspitze zeigen *keine* Kausaleinflüsse ein (was in DAGs nicht erlaubt wäre).
- Stattdessen zeigen Sie eine Assoziation bedingt durch eine (nicht aufgeführte) Konfundierungsvariable an.

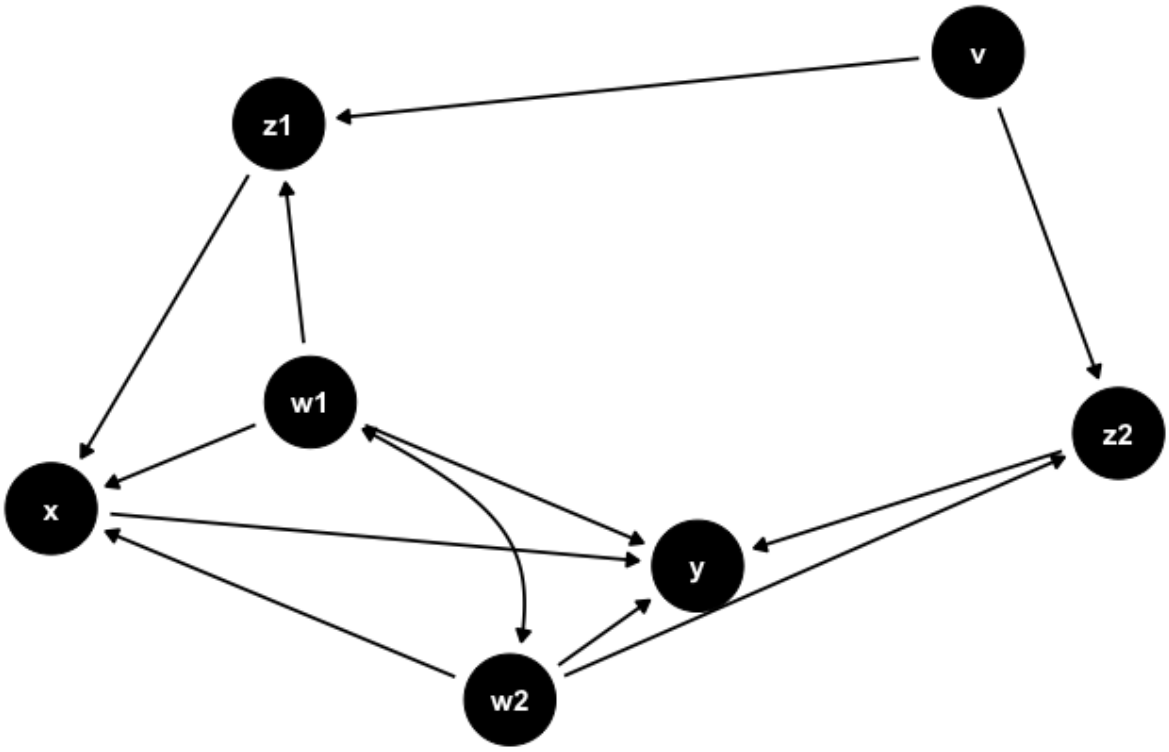
a. { w2, z2 }
b. { w1 }
c. { w1, w2, z2 }
d. { w1, z2 }
e. { w1, w2 }

Lösung

```
{ w1, w2, z2 }
{ v, w1, w2 }
{ w1, w2, z1 }
```

a. Falsch
b. Falsch Die Regressionsformel lautet also. `y ~ x + w1 + w2 + z2`. Es gibt mehrere *Adjustment Sets*, aber unsere Lösungsoptionen lassen nur eine zu. Alternative Visualisierung: `text dag <- dagitty::dagitty( "dag { y <- x <- z1 <- v -> z2 -> y z1 <- w1 <-> w2 -> z2 x <- w1 -> y`

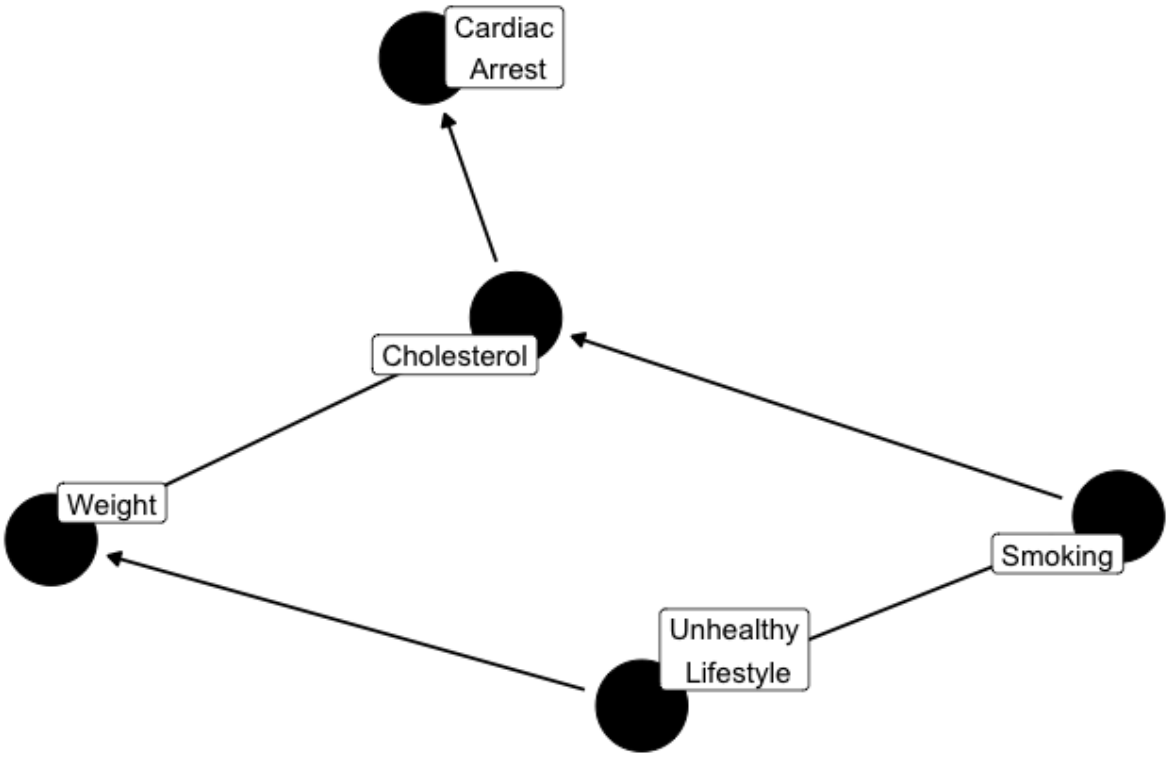
```
x <- w2 -> y x [exposure] y [outcome] }" ) ggdag(dag) + theme_dag()
```



c. Wahr
d. Falsch
e. Falsch

6. **Aufgabe**

Gegeben sei ein DAG `g` (s.u.). Was ist die minimale Menge an Variablen (minimal adjustment set), die man kontrollieren muss, um den kausalen Effekt von `smoking` auf `arrest` zu identifizieren?

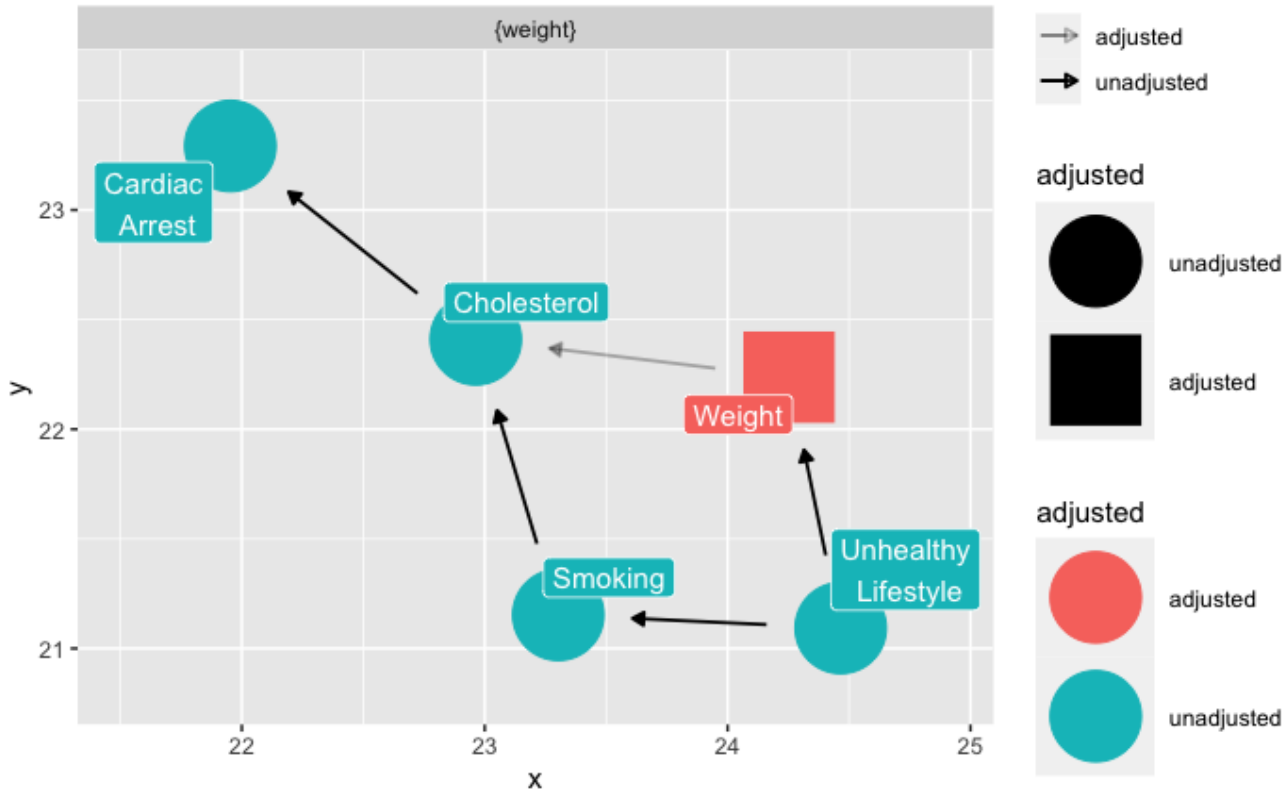


a. keine, da nicht identifiziferbar
b. { Cholestrol, Unhealty Lifestyle }
c. { Cholestorol }
d. { Weight }
e. { Cholestorol, Weight }

Lösung

```
{ weight }
```

Durck die Kontrolle von `weight` wird der gesuchte kausale Effekt identifizierbar.



Also lautet die Regressionsformel: `arrest ~ smoking + weight`.

Es wäre ein fataler Fehler, nähme man den Mediator `Cholestorol` mit auf:

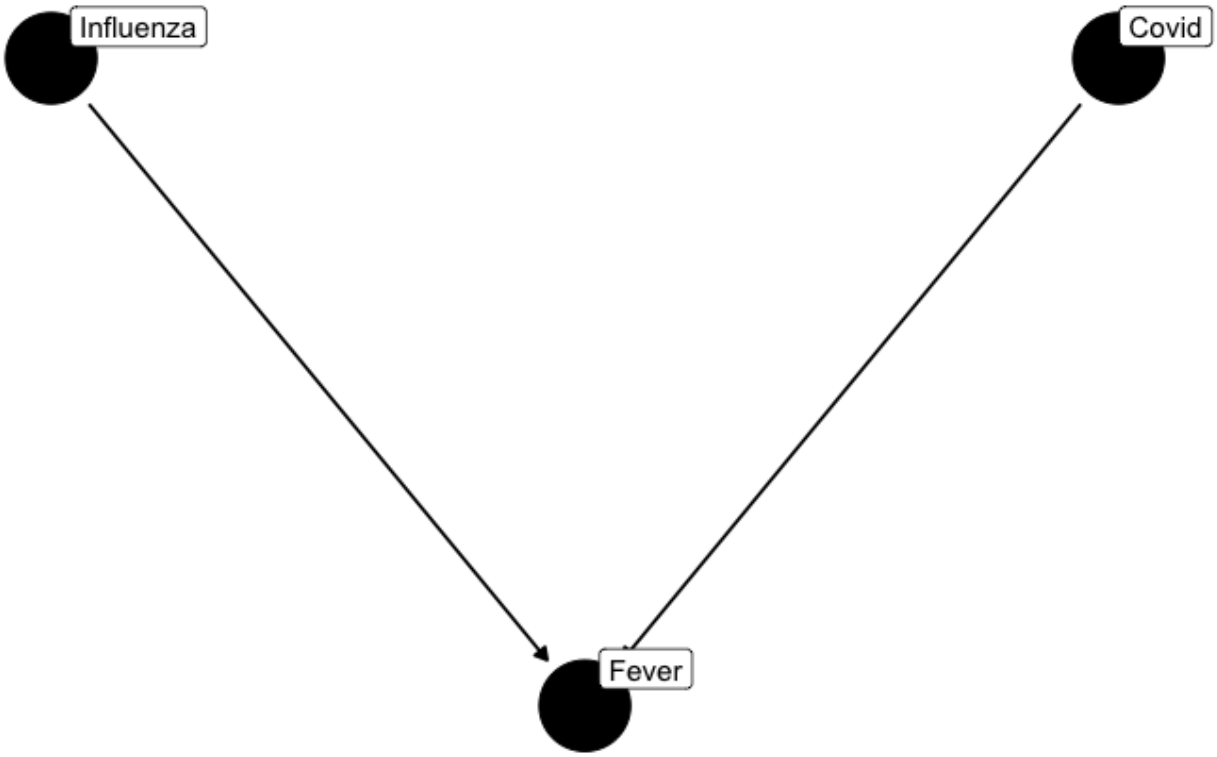
`arrest ~ smoking + cholestorol.` 🌩️🛑

a. Falsch
b. Falsch
c. Falsch
d. Wahr
e. Falsch

7. **Aufgabe**

Im Rahmen einer Studie soll untersucht werden, ob eine Influenza-Infektion einen (kausalen) Einfluss auf eine Covid19-Infektion hat.

In Wahrheit (aber unbekannt) sei der DAG wie folgt (s.u.).

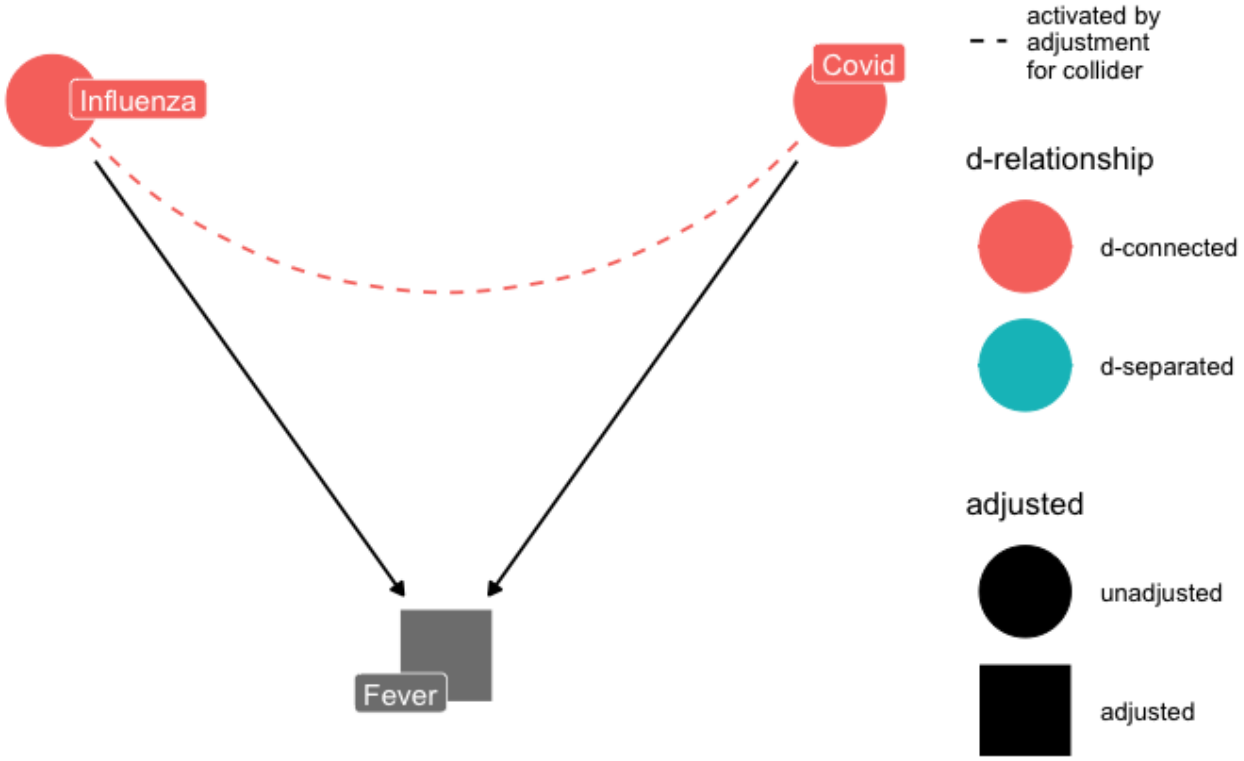


Ist es sinnvoll, das Auftreten von *Fieber* (`Fever`) zu kontrollieren?

a. Ja, durch eine Kontrolle von `Fever` ist ein kausaler Effekt identifizierbar
b. Nein, da eine Kontrolle von `Fever` eine Verzerrung erzeugt wird (Konfundierung)
c. Nein, da durch eine Kontrolle von `Fever` eine Verzerrung erzeugt wird (Kollisionsverzerrung)
d. Nein, da eine Kontrolle von `Fever` nicht nötig ist (aber auch nicht schädlich)
e. Ja, eine Kontrolle von `Fever` ist zwar nicht nötig, aber wird zu exakteren Ergebnissen führen

Lösung

Nein, da durch eine Kontrolle von `Fever` eine Verzerrung erzeugt wird (Kollisionsverzerrung): `Influenza` und `Covid` sind dann `d-connected`, obwohl es in Wirklichkeit *keinen* kausalen Pfad zwischen den beiden Variablen gibt.

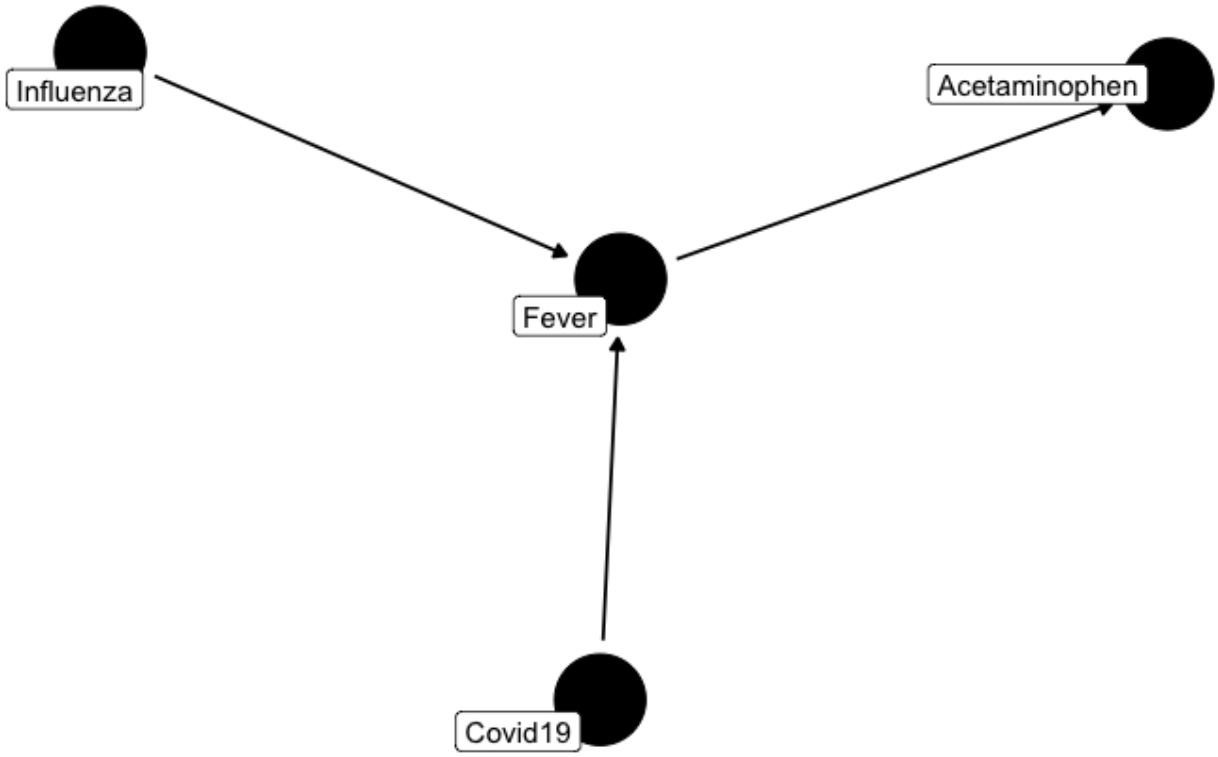


a. Falsch
b. Falsch
c. Wahr
d. Falsch
e. Falsch

8. **Aufgabe**

Im Rahmen einer Studie soll untersucht werden, ob eine Influenza-Infektion einen (kausalen) Einfluss auf eine Covid19-Infektion hat. Außerdem wird dabei der Nutzen des Medikaments Acetaminophen untersucht.

In Wahrheit (aber unbekannt) sei der DAG wie folgt (s.u.).



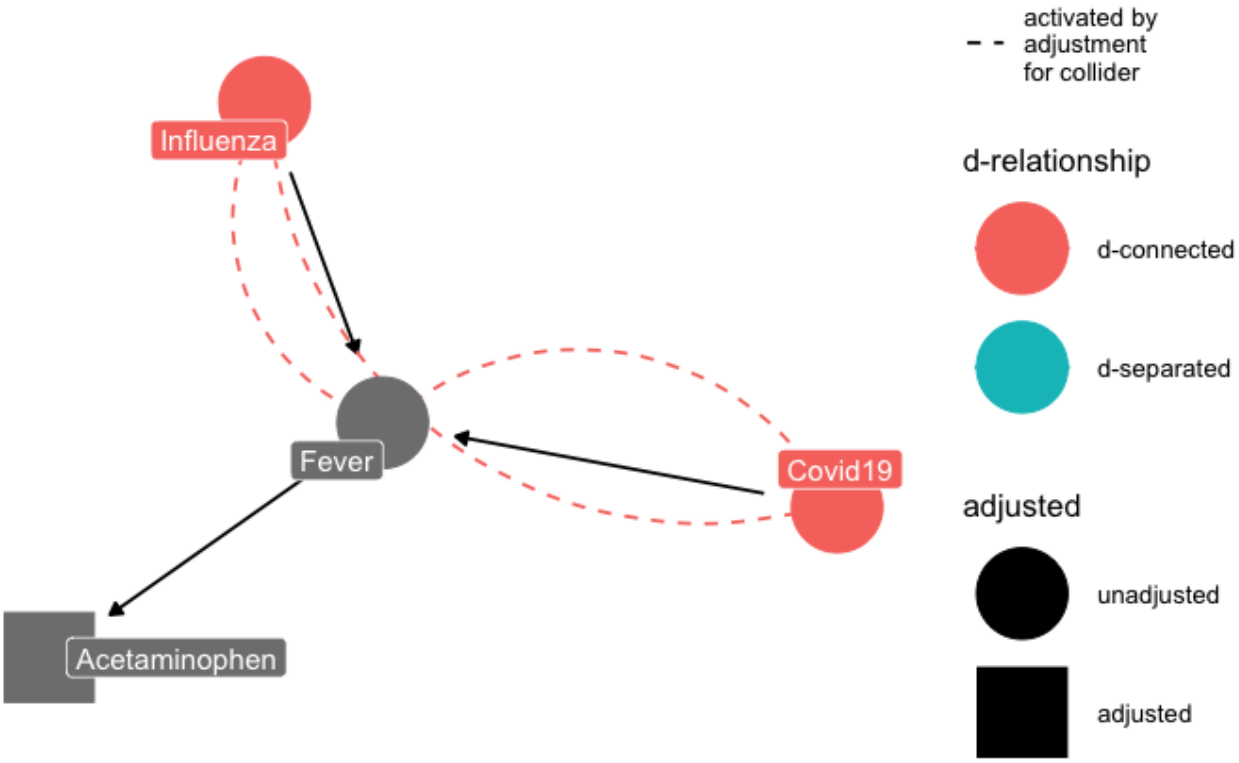
Ist es sinnvoll, die Einnahme von Fiebersenker (*Acetaminophen*) zu kontrollieren?

a. Nein, es ist nicht sinnvoll, da es nicht nötig ist (aber auch nicht schädlich)
b. Ja, nur so ist ein kausaler Effekt identifizierbar
c. Nein, es ist nicht sinnvoll,da durch eine Kontrolle von Acetaminophen eine Verzerrung erzeugt wird (Konfundierung)
d. Nein, es ist nicht sinnvoll, da durch eine Kontrolle von Acetaminophen eine Verzerrung erzeugt wird (Kollision)
e. Ja, es ist nicht nötig, aber wird zu exakteren Ergebnissen führen

Lösung

Nein, es ist nicht sinnvoll zu kontrollieren, da eine Verzerrung erzeugt wird (Kollision).

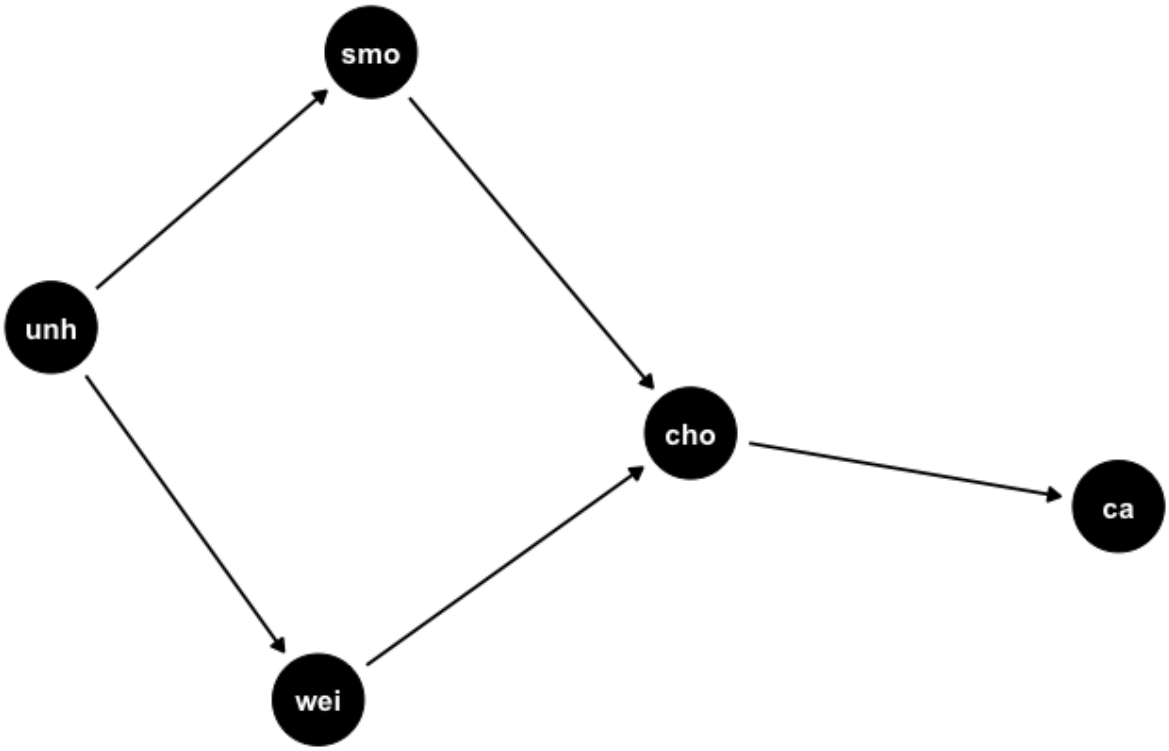
Acetaminophen ist ein Nachfahre von `Fever`, also wird die Kontrolle dieser Variable den grundsätzlich gleichen (nur etwas schwächeren) Effekt haben, wie das Kontrollieren der “kausalen Vorfahren”, hier `Fever`.



a. Falsch
b. Falsch
c. Falsch
d. Wahr
e. Falsch

9. **Aufgabe**

Eine Forscherin untersucht den Zusammenhang von Rauchen `smo` (smoking, UV, exposure) und Herzstillstand `ca` (cardiac arrest, AV, outcome). Sie hegt die Hypothese, dass Rauchen einen Einfluss auf den Cholesterolspiegel `cho` (cholestorol) hat, was wiederum Herzstillstand auslösen könnte.



Hier sehen Sie die Definition des DAGs:

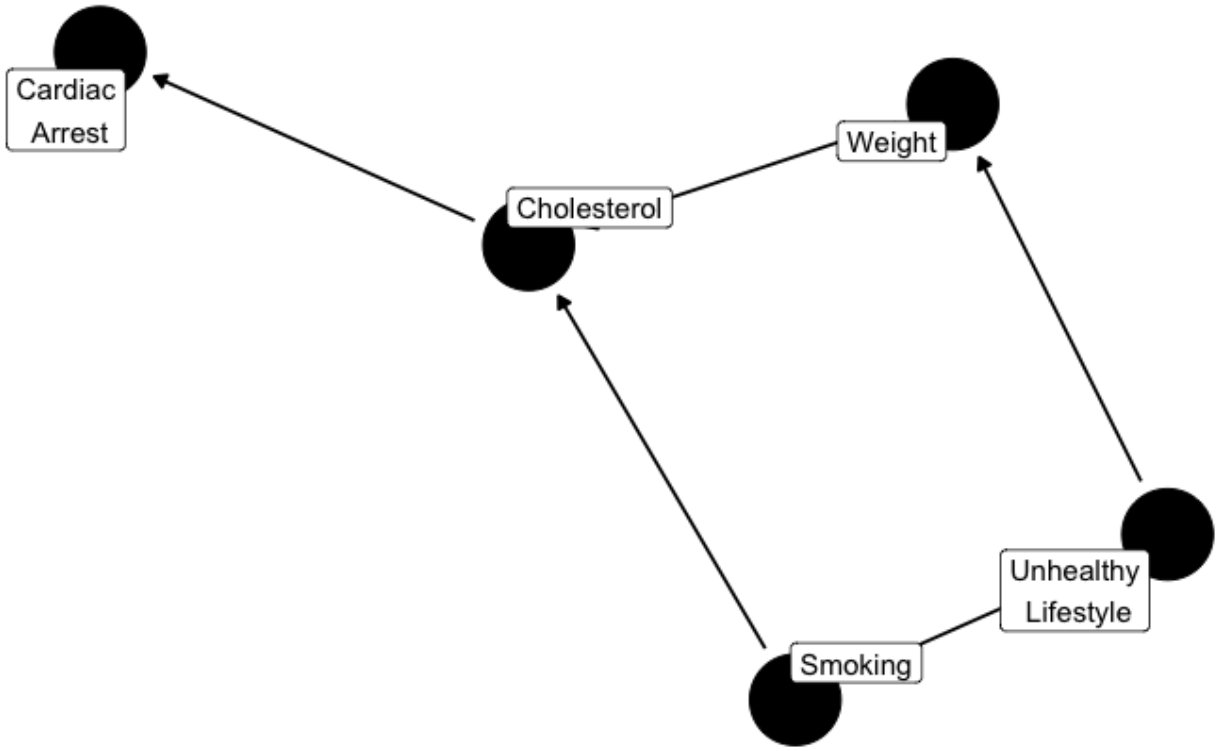
```
## dag {
## ca [outcome]
## cho
## smo [exposure]
## unh
## wei
## cho -> ca
## smo -> cho
## unh -> smo
## unh -> wei
## wei -> cho
## }
```

Die Forscherin überlegt, `Cholestorol` zu kontrollieren. Ist diese Idee sinnvoll?

a. Nein, da eine Kollision erzeugt wird.
b. Es schadet nicht, aber es ist auch nicht nötig.
c. Ja, *nur* so wird der kausale Effekt identifiziert.
d. Ja, so wird der kausale Effekt identifiziert.
e. Nein, da die Assoziation zwischen UV und AV unterbrochen wird.

Lösung

Alternative Visualisierung:



Nein, es ist nicht sinnvoll, da die Assoziation zwischen UV und AV unterbrochen wird. Damit wird der Kausaleffekt von Rauchen auf den Herzstillstand “wegkontrolliert”. Die Ergebnisse würden dann fälschlich aufzeigen, dass Rauchen nicht in Verbindung stünde mit Herzstillstand, was falsch ist.

```
## Error in .checkAllNames(x, value): cholesterol is not a variable in `x`
```

Stattdessen wäre es nötig, `weight` oder auch `unhealthy lifestyle` zu kontrollieren, um den kausalen Effekt von `smoking` auf `cardiacarrest` zu identifizieren.

Hier sind die möglichen “Adjustment Sets”, die Mengen der Variablen, die man (pro Menge) kontrollieren muss, um den gesuchten Kausaleffekt zu identifizieren:

```
## Error in .checkAllNames(x, c(exposure, outcome)): smoking is not a variable in `x`
```

a. Falsch
b. Falsch
c. Falsch
d. Falsch
e. Wahr

10. **Aufgabe**

Betrachten wir den Datensatz `SaratogaHouses`, den Sie hier herunterladen können. Ein Codebook findet sich hier.

Sie kommen auch so an die Daten ran:

```
library(mosaicData)
data("SaratogaHouses")
```

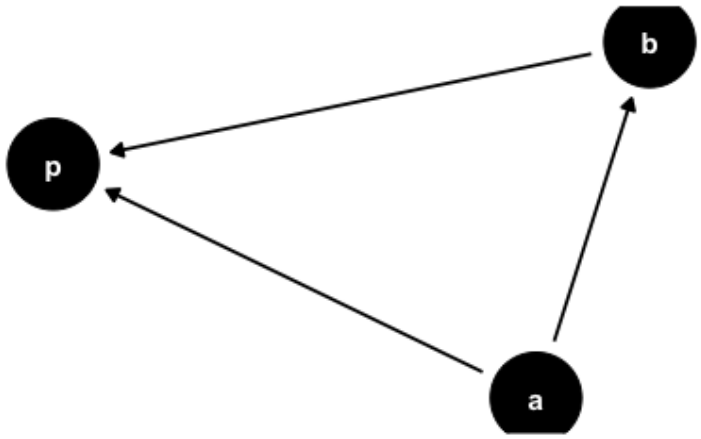
Gegeben sei in diesem Zusammenhang folgender DAG:

```
dag1 <- "
dag{
a -> p
a -> b -> p
}
"
```

Wobei `a` für *(living) area* steht, also der Wohnfläche eines Hauses, `b` für *bedrooms*, der Anzahl der Schlafzimmer und `p` für *prize*, den Preis, den das Haus beim Verkauf erzielt hat.

So sieht das dann aus:

```
ggdag(dag1) + theme_dag()
```



UV sei `a`; AV sei `p`.

a. Berechnen Sie den *direkten* Effekt der *Wohnfläche* auf den *Preis*!
b. Berechnen Sie den *totalen* Effekt der Wohnfläche auf den Preis!

Hinweise: - Mit *direkter* Effekt ist der kausale Effekt von UV auf AV - ohne Zwischenglieder (Mediatoren) - gemeint. - Mit *indirekter* Effekt ist der kausale Effekt von UV über einen (oder ggf. mehrere) Mediator(en) auf die AV gemeint. - Mit *totaler* Effekt ist die Summe des direkten plus des oder der indirekten Effekte gemeint. - Geben Sie jeweils den Punktschätzer eines linearen Regressionsmodells an! - Gehen Sie vom oben genannten DAG aus. - Runden Sie ohne Dezimalstellen.

Lösung

```
d <-
  SaratogaHouses %>%
  select(price, bedrooms, livingArea) %>%
  drop_na()
```

a. direkter Effekt:

```
direkter_eff_lm <-
  stan_glm(price ~ bedrooms + livingArea,
           data = d,
           refresh = 0)
coef(direkter_eff_lm)
```

```
## (Intercept)    bedrooms  livingArea
##  36473.9406 -14159.9736    125.4215
```

Um einen direkten Effekt zu berechnen, müssen wir den *spezifischen*, uniquen Effekt der UV berechnen. Das erreichen wir durch eine multiple Regression, in der also die übrigen Prädiktoren aufgenommen sind. Das Resultat ist ein Koeffizient für die Assoziation der UV mit der AV, bereinigt um die Zusammenhänge der übrigen Prädiktoren.

Zur Erinnerung: Die multiple Regression liefert Koeffizienten pro Prädiktor, die bereinigt sind um den (statistischen) Einfluss der anderen Prädiktoren, mit anderne Worten: die Koeffizienten der multiplen Regression zeigen den Effekt von “nur diesem Prädiktor”.

Der Punktschätzer für den direkten Effekt (von Wohnfläche) ist:

```
direkter_eff <-
  coef(direkter_eff_lm)[3] %>%
  round(0)

direkter_eff
```

```
## livingArea
##        125
```

b. totaler Effekt:

```
## (Intercept)  livingArea
##  13275.7939    113.2647
```

Der totale Effekt lässt sich berechnen, in dem man keine weiteren Prädiktoren neben der UV in die Regression mitaufnimmt. Die *einfache* (univariate) Regression zeigt den totalen Effekt der UV auf die AV.

Der Punktschätzer für den totalen Effekt beträgt:

```
## livingArea
##        113
```