Observational causal inference

The same research team that led the randomized controlled trial of a technical training program in Ecuador then rolled out the program to Guatemala. However, they did not have funding to run a complete RCT. Instead, they allowed anyone to sign up for the program.

Your colleague attempted to measure the causal effect of this training program on incomes. They again conducted some statistical analysis in R, but again, they omitted all explanation and interpretation. They've moved and you don't know their new contact information.

You have access to the original data, which contains these columns:

Variable name	Description
`id`	Person's ID number
`wage`	Monthly income before intervention (in USD)
`training`	Indicator for whether the person participated in training program
`age`	Person's age
`education`	Person's education (in years)
`computer`	Person's knowledge of computers, on a 0-10 scale
`internet`	Indicator for whether the person has regular internet access
`internet`	Indicator for whether the person heard about the program beforehand

Given the information provided below, interpret the results from this analysis, as well as any assumption checks or tests your colleague included. Did this program have have an effect on wages? How much? Is it significant?

```r
library(tidyverse)
library(broom)
library(ggdag)
library(dagitty)

training_data <- read_csv("training.csv")
```

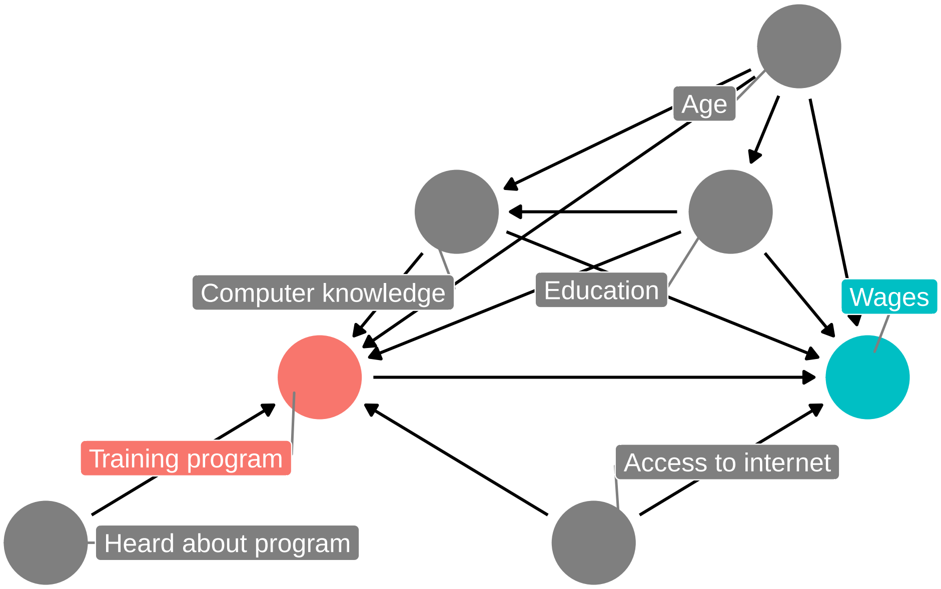
```r
head(training_data)
```

id	wage	training	age	education	computer	internet	heard
1	1555	TRUE	21	11	3	FALSE	FALSE
2	1550	FALSE	29	17	7	FALSE	FALSE
3	1437	FALSE	37	18	5	FALSE	TRUE
4	881	TRUE	19	10	3	FALSE	FALSE
5	1441	FALSE	30	15	5	TRUE	FALSE
6	1262	TRUE	31	16	4	TRUE	FALSE

↓ 1: What's going on here? ↓

```
training_dag <- dagify(wage ~ training + age + education + computer + internet,
                       training ~ age + education + computer + internet + heard,
                       computer ~ education + age,
                       education ~ age,
                       exposure = "training",
                       outcome = "wage",
                       labels = c("wage" = "Wages", "training" = "Training program",
                                  "age" = "Age", "education" = "Education",
                                  "computer" = "Computer knowledge",
                                  "internet" = "Access to internet",
                                  "heard" = "Heard about program"),
                       coords = list(x = c(wage = 4, training = 2, age = 3.75, education = 3.5,
                                           computer = 2.5, internet = 3, heard = 1),
                                     y = c(wage = 2, training = 2, age = 4, education = 3,
                                           computer = 3, internet = 1, heard = 1)))

ggdag_status(training_dag, text = FALSE, use_labels = "label", seed = 123) +
  guides(color = FALSE) +  # Turn off legend
  theme_dag()
```

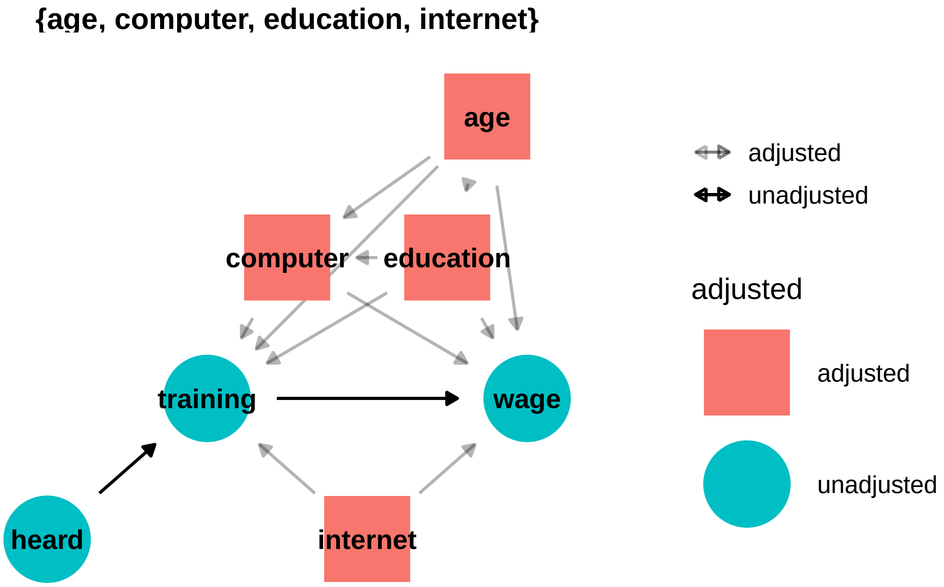


↓ 2: What's going on here? ↓

```
adjustmentSets(training_dag)
```

```
## { age, computer, education, internet }
```

```
ggdag_adjustment_set(training_dag, shadow = TRUE, text_col = "black") +
  theme_dag()
```



↓ 3: What's going on here? ↓

```r
model_predict_training <- glm(training ~ age + computer + education + internet,
                              data = training_data,
                              family = binomial(link = "logit"))

training_adjusted <- augment_columns(model_predict_training, training_data,
                                     type.predict = "response") %>%
  rename(propensity = .fitted) %>%
  mutate(ipw = (training / propensity) + ((1 - training) / (1 - propensity)))

training_adjusted %>%
  select(id, wage, training, propensity, ipw) %>%
  head()
```

id	wage	training	propensity	ipw
1	1555	TRUE	0.725	1.38
2	1550	FALSE	0.517	2.07
3	1437	FALSE	0.326	1.48
4	881	TRUE	0.768	1.30
5	1441	FALSE	0.666	2.99
6	1262	TRUE	0.623	1.60

↓ 4: What's going on here? ↓

```r
model_ate <- lm(wage ~ training, data = training_adjusted, weights = ipw)
tidy(model_ate)
```

term	estimate	std.error	statistic	p.value
(Intercept)	1219	9.62	126.76	0
trainingTRUE	84	13.58	6.18	0