Lecture 20

Application of randomized experiments to workplace wellness

Julian Reif
Fall 2025

RStudio setup for this lecture

- Log into RStudio on your Amazon EC2 instance
  - Use AMI `FIN550-RStudio` with IAM role `BigDataEC2Role`

```
# This is a Unix command. Enter via RStudio Terminal
aws s3 cp --recursive s3://bigdata-fin550-reif/lecture-20 ~/fin550/lecture-20
```

Workplace wellness programs are growing rapidly

- Workplace wellness is an $18 billion industry in the United States
  - Aims: reduce health care costs, improve employee health, increase productivity
  - 83% of large firms offer wellness programs, covering over 50 million workers

- Workplace wellness is popular among policymakers
  - US legislation encourages these programs
  - Some advocate expanding to the Medicare and Medicaid insurance programs

- Workplace wellness programs are controversial
  - Do they actually improve health and productivity?

Prior evidence is limited

- Meta-analysis by Baicker, Cutler, and Song (2010) found:
  - Medical cost reduction of $3.27 for every $1 spent on wellness
  - Absenteeism cost reduction of $2.73 for every $1 spent on wellness

- Empirical challenges:
  - Selection bias from non-random participation
  - Measuring outcomes
  - Statistical power

The Illinois Workplace Wellness Study

- Two-year randomized experiment at University of Illinois
  - 12,459 employees invited: faculty (26%), administrative and union positions (74%)
  - Individual, random assignment to control or treatment groups

- Rich data linked at individual level allows for comprehensive evaluation
  - Administrative data on employment and health insurance claims
  - Administrative data on health behaviors (gym use, running events)
  - Survey data
  - Detailed biometric data

Study design

Background on workplace wellness programs

- Three main components:
  1. Biometric health screening
  2. Health risk assessment (HRA)
  3. Wellness activities

- We designed a "gold-standard" wellness program (iThrive)
  - Includes all three components above
  - Includes financial incentives tied to participation
  - Allows employees to take paid time off to participate

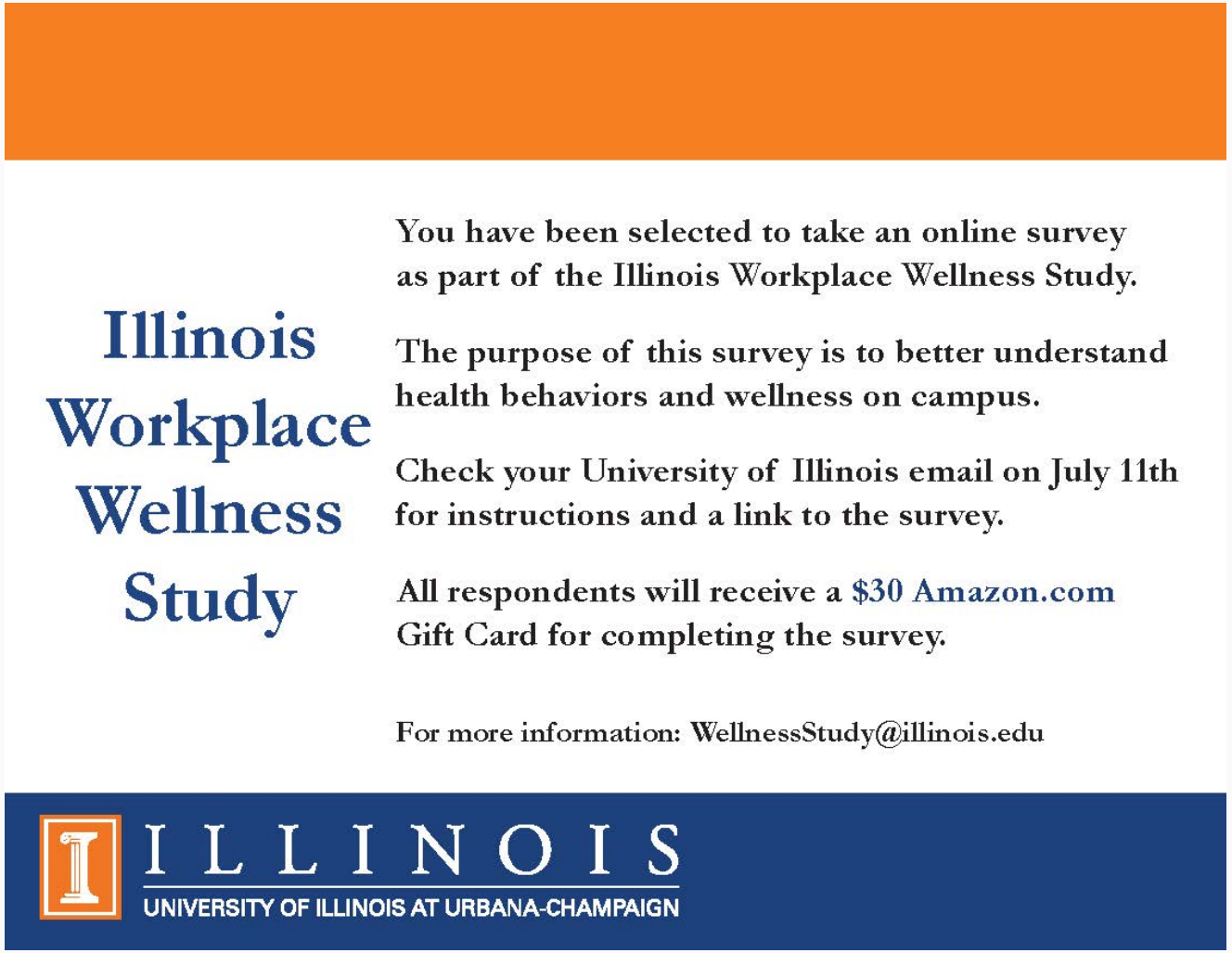
Study enrollment (July 2016)

- Enrollment required completing a 15-minute online survey

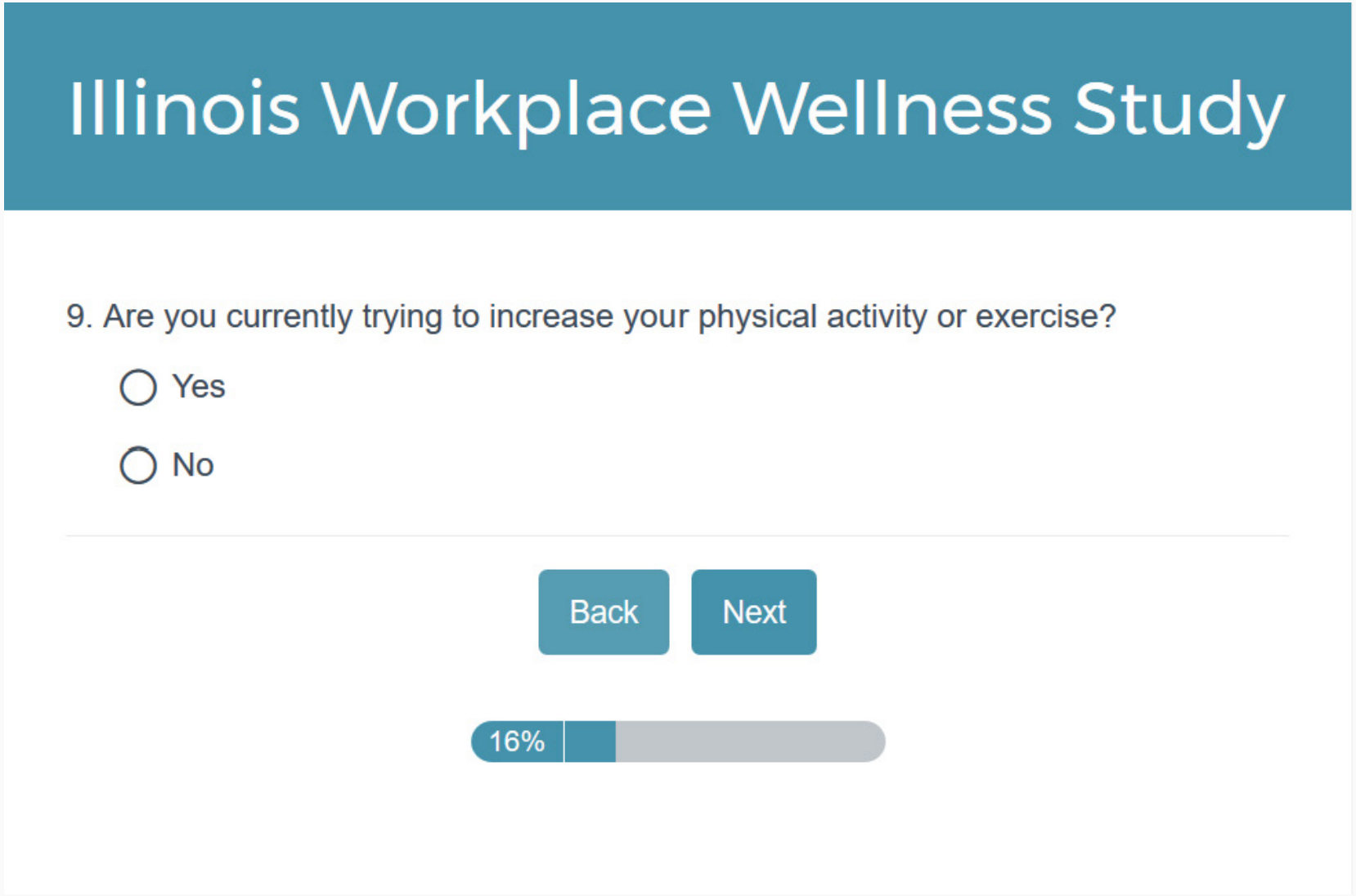
- Survey invitations sent to 12,459 employees
  - Postcard notification sent to employee home address (July 6, 2016)
  - Email invitation to employees, with personalized link to online survey (July 11)

- Employees were offered $30 Amazon.com gift card, plus a chance "to participate in a second part of the research study"

Survey invitation



Online survey interface: sample question



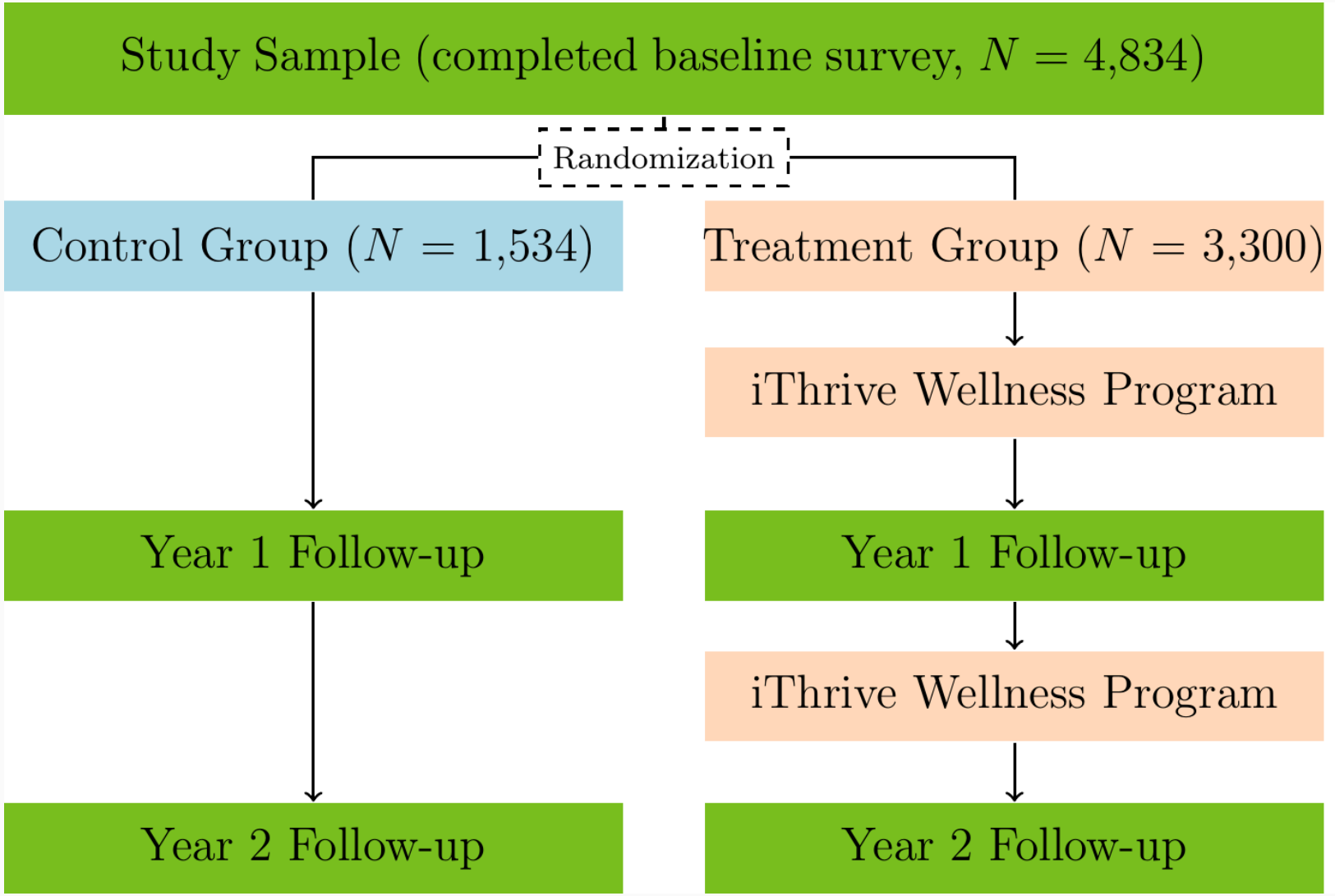
Survey responses were very complete

- **Response:** 4,834 employees (38.8%) successfully completed baseline survey
  - Survey open for three weeks, with periodic email reminders

- **Completeness:** Fewer than 1% of respondents skipped *any* question

- **Validity:** Age, sex responses match University administrative records closely
  - Last questions on the survey
  - 99.4% within 1 year of age reported in University records

Experimental design



Participating in iThrive involved two steps

Step 1: Screening: up to $200 reward

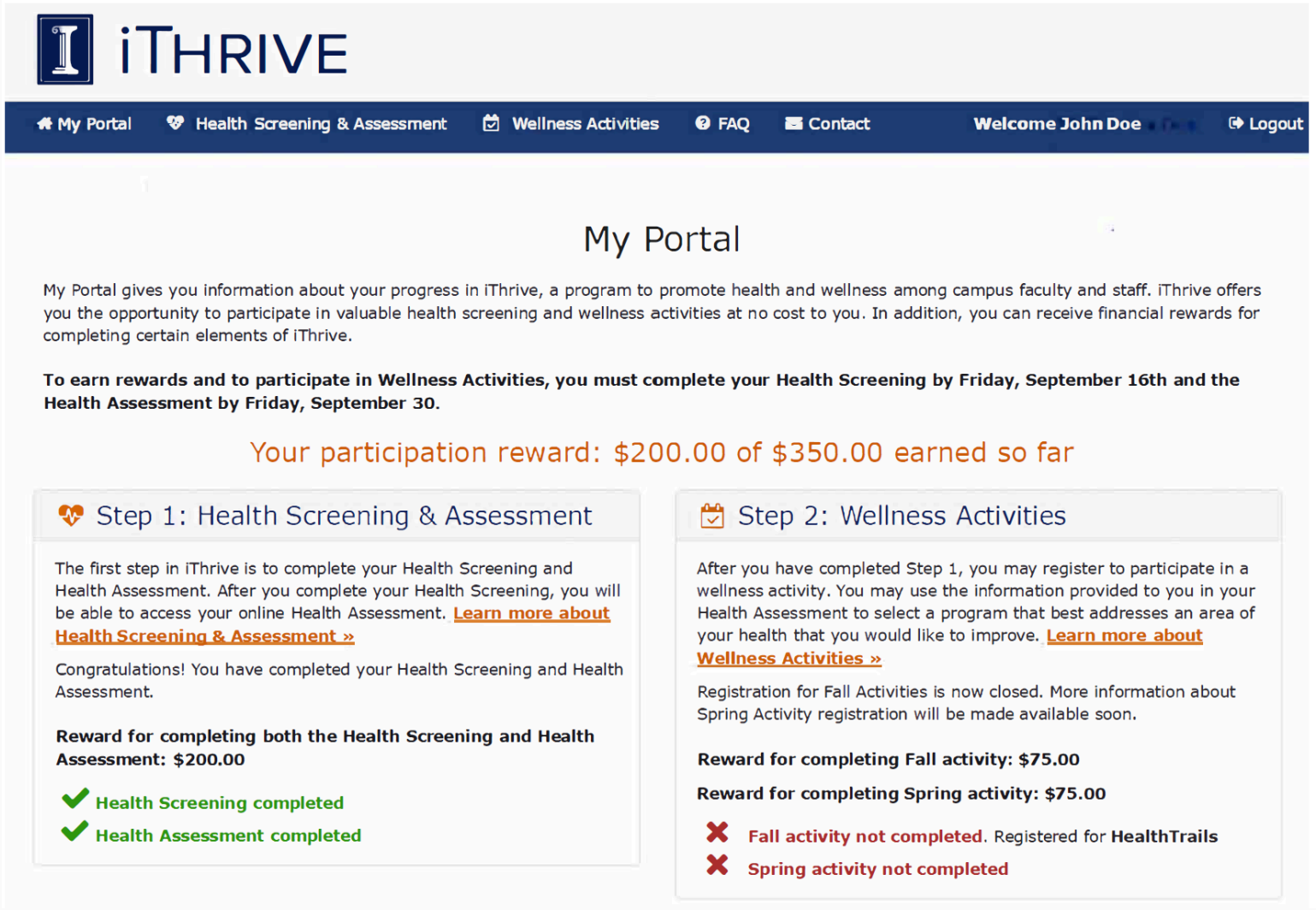
- *Biometric Screening*
  - 8 different on- and off-campus locations
  - Fingerstick + blood pressure, height, weight, waist circumference

- *Health Risk Assessment (HRA)*
  - Online questionnaire designed to assess lifestyle habits

Step 2: Wellness Activities: up to $75 per semester

- Many options, e.g., Weight Watchers, smoking cessation, stress management
- Classes ranged from 6-12 weeks
- Completion defined as attending at least 3/4 of sessions

Website portal



Data and results

We constructed 42 outcomes from our datasets

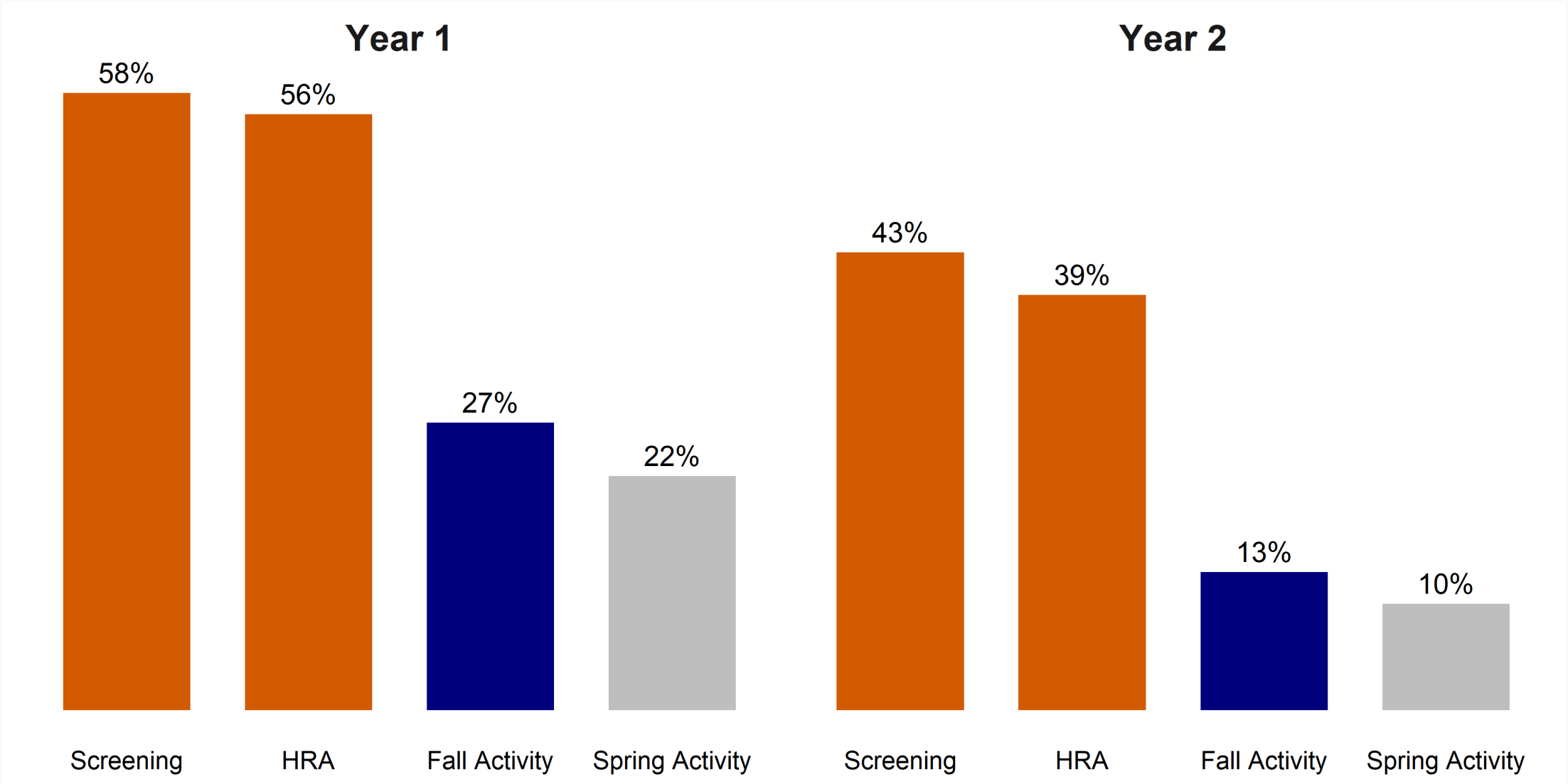
- Medical spending and utilization (8 outcomes)
- Health status and behaviors (17 outcomes)
- Employment and productivity (17 outcomes)

- Today: report results for a few focal outcomes

Balance table (control and treatment group means)

	Control	Treat	p-value (difference in means)
Male	0.426	0.428	0.90
Age 50+	0.323	0.327	0.82
White	0.841	0.836	0.65
Faculty	0.196	0.201	0.72
Total spending (dollars/month)	506	465	0.32
Running event participant	0.107	0.118	0.13
Gym visits (days/year)	7.36	6.78	0.48
Sick leave (days/year)	6.05	6.13	0.71
Annual salary (dollars)	61,528	61,736	0.84
Sample size	1,534	3,300	

Treatment group participation in the program



We present two sets of results

- Selection into the workplace wellness program
  - Treatment group is *invited* to participate in the program
  - $\rightarrow$ Not everybody in the treatment group chooses to participate
    - Is decision to participate as good as random?
    - Or is there selection bias?

- Causal effects of the workplace wellness program
  - Short-run (12 months)
  - Longer-run (24-30 months)

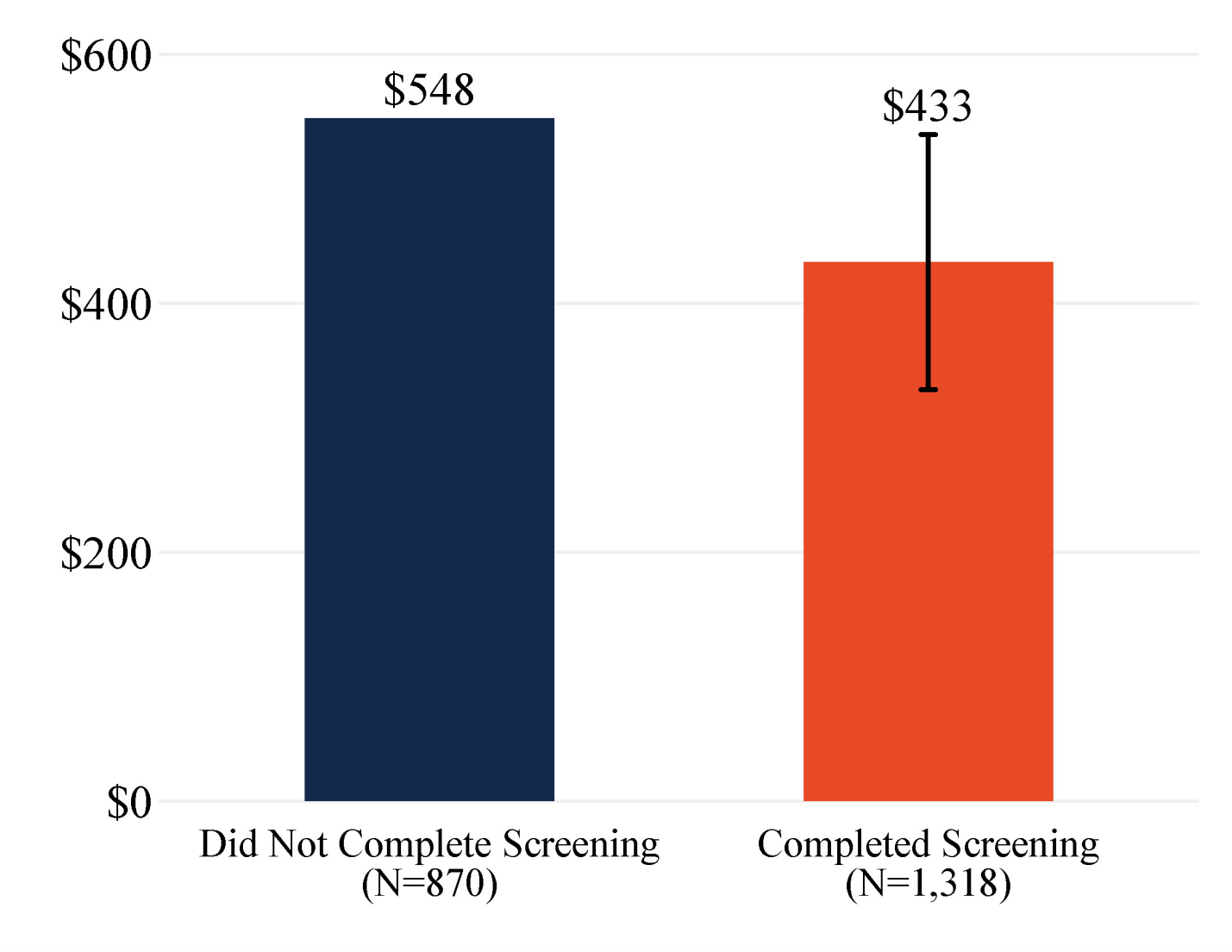
Results 1: Who participates?

$$X_i = \alpha + \theta P_i + \varepsilon_i$$

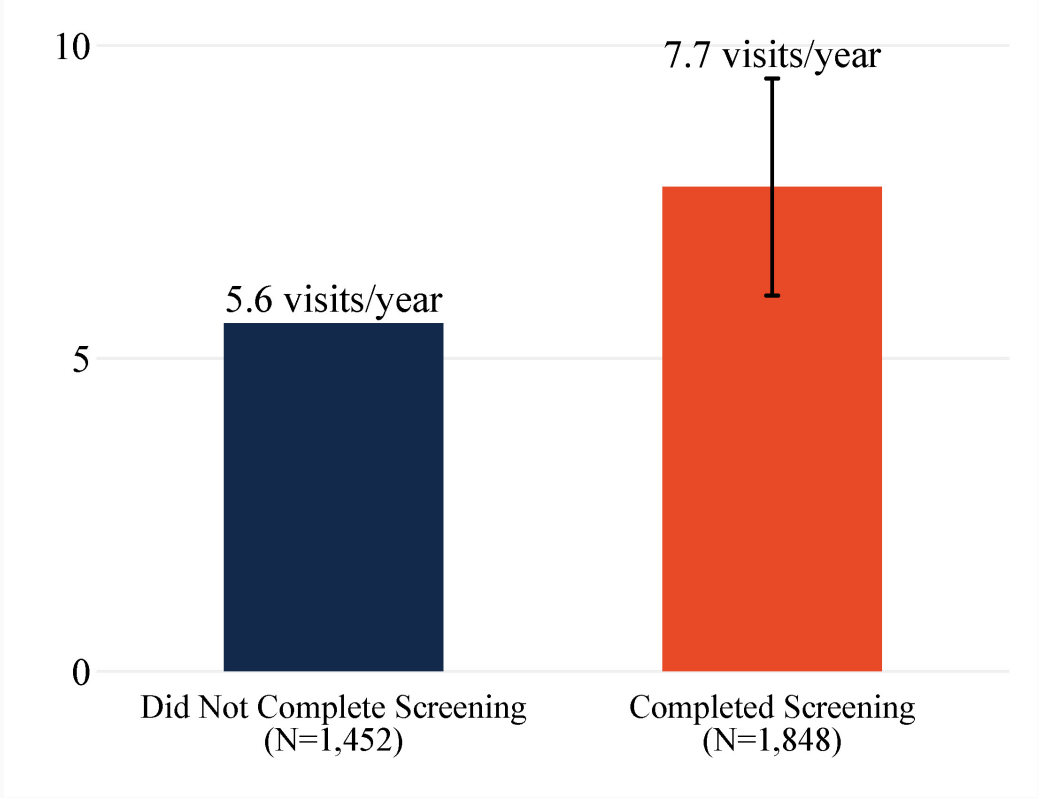
- Observations: employees assigned to the treatment group $(N = 3,300)$
- X_i: characteristic of employee i, **measured prior to the study**
- P_i: dummy variable for completing both health screening and HRA in the first year
  - θ measures a difference in means of X: avg(completed) $-$ avg(did not complete)

- If decision to participate is as good as random, what should the coefficient θ be?
- What does it mean if we find that the coefficient $\theta \neq 0$?

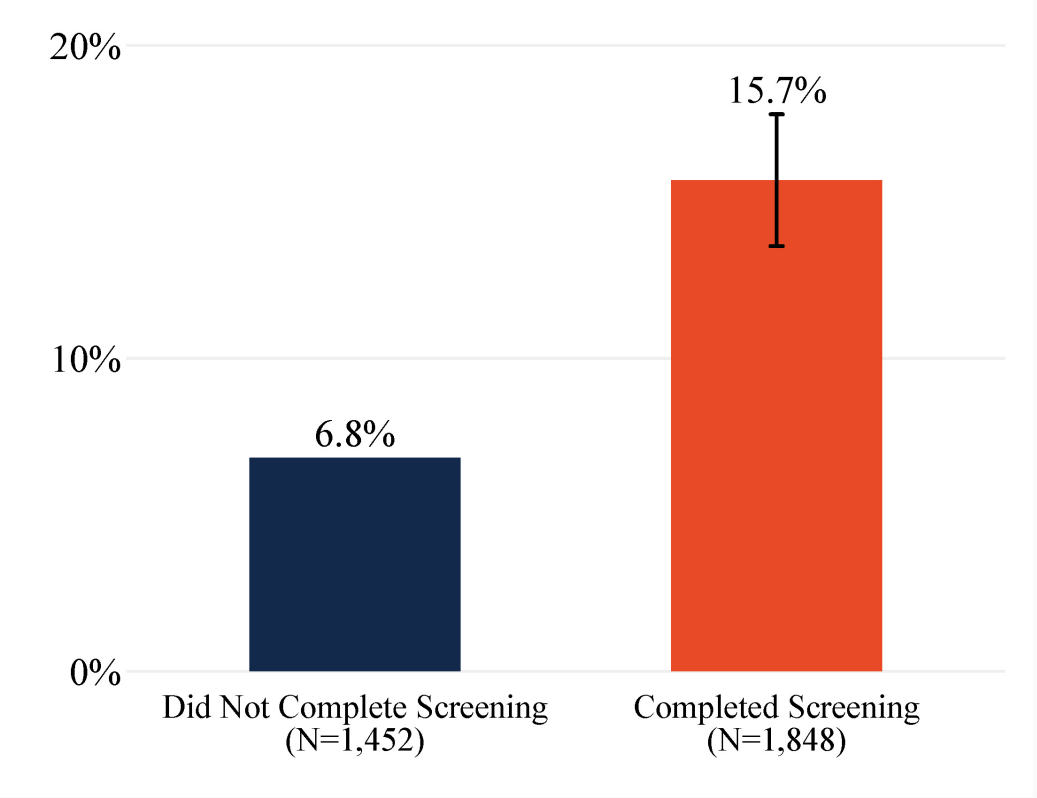
Participants already had lower health spending



Participants already had healthier behaviors



Average annual gym visits



Running event participation

Results 2: Causal effects of workplace wellness

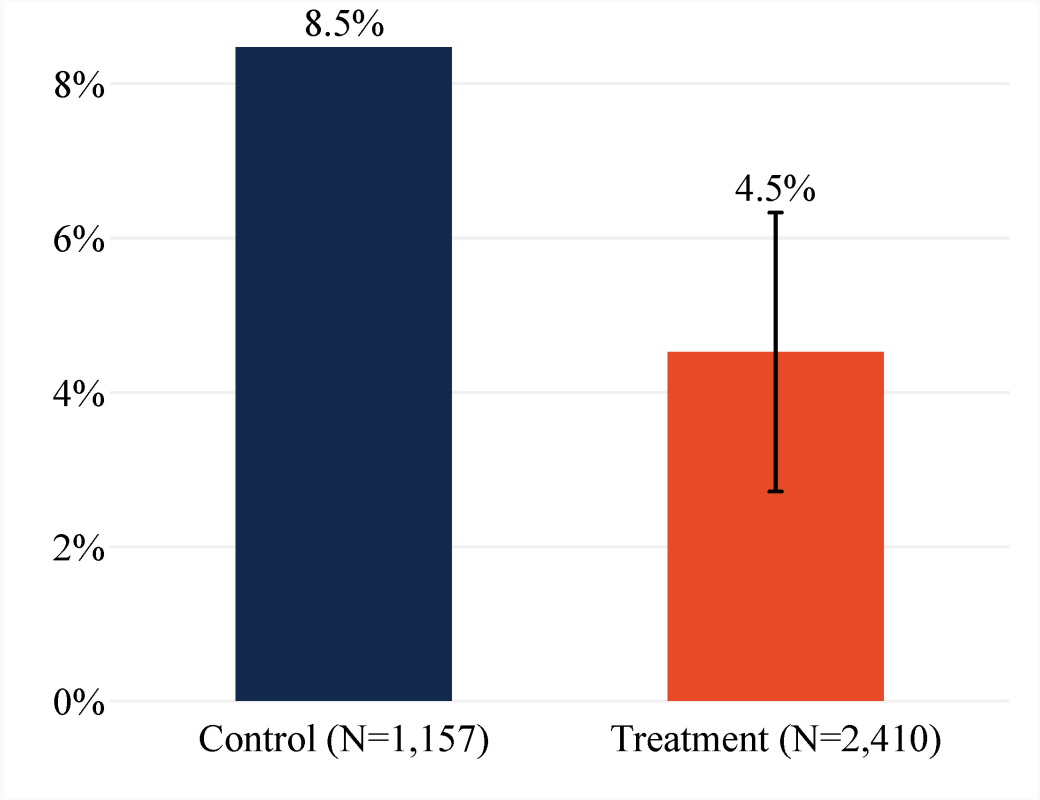
$$Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i$$

- Observations: employees assigned to treatment or control group $(N = 4,834)$

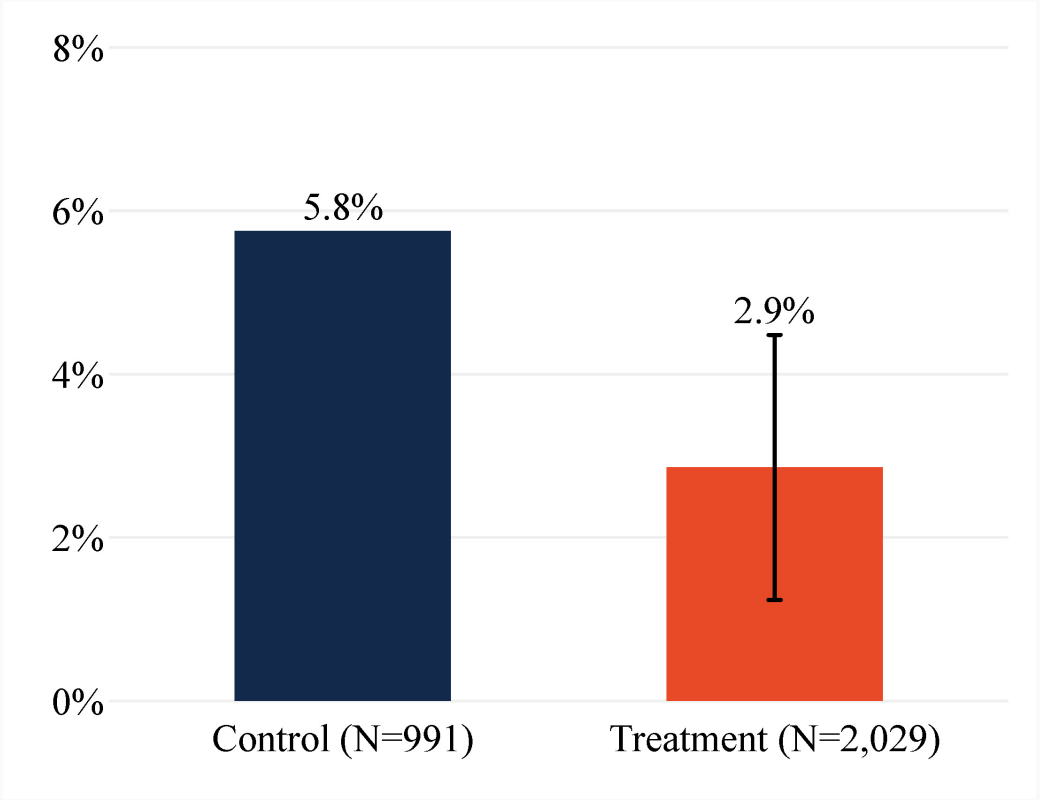
- Y_i: outcome variable
  - We consider 42 different survey and administrative data outcomes

- T_i: dummy for treatment assignment (1 = treatment, 0 = control)

Program reduced number of people with no screenings

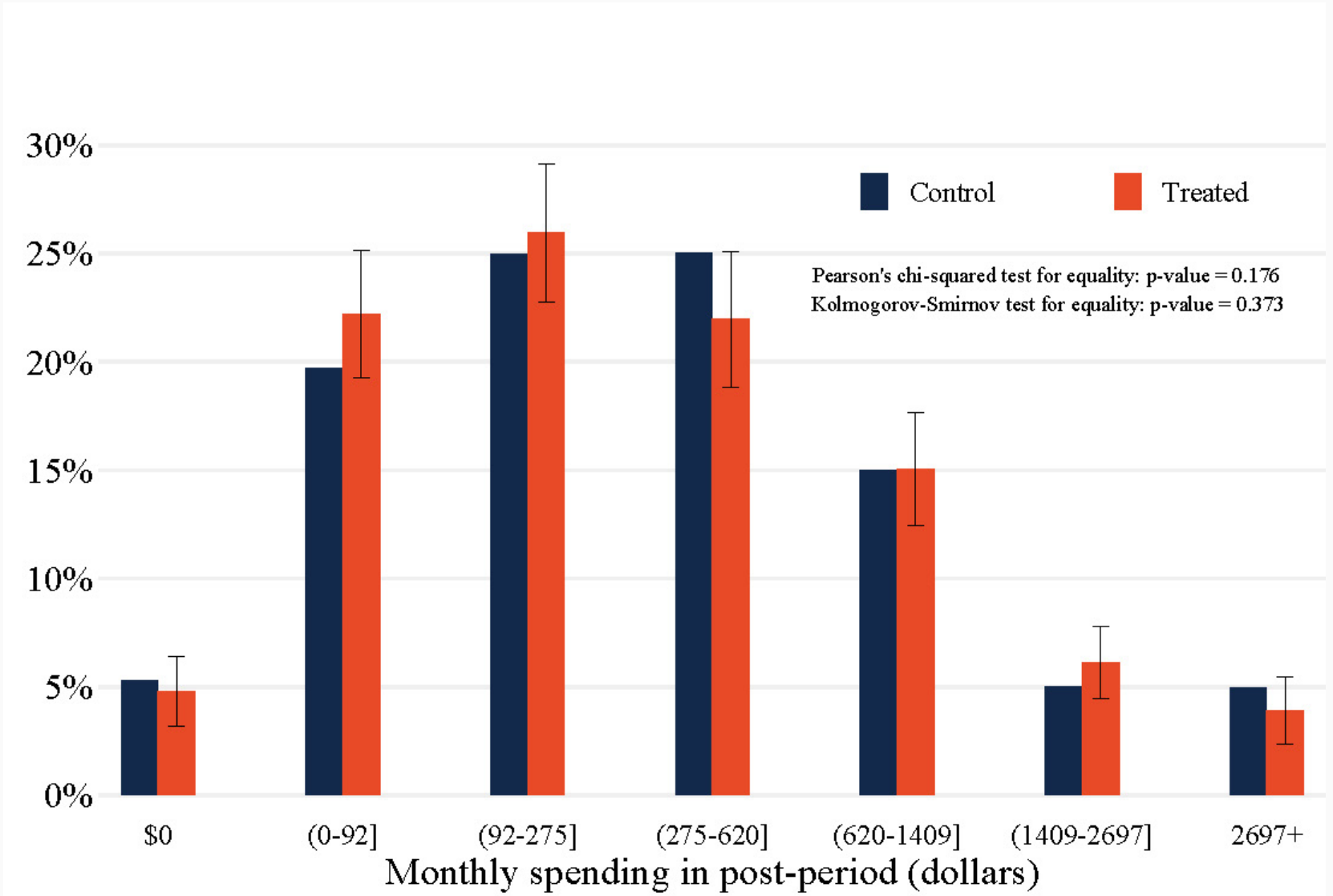


Never had a screening (12-month survey)

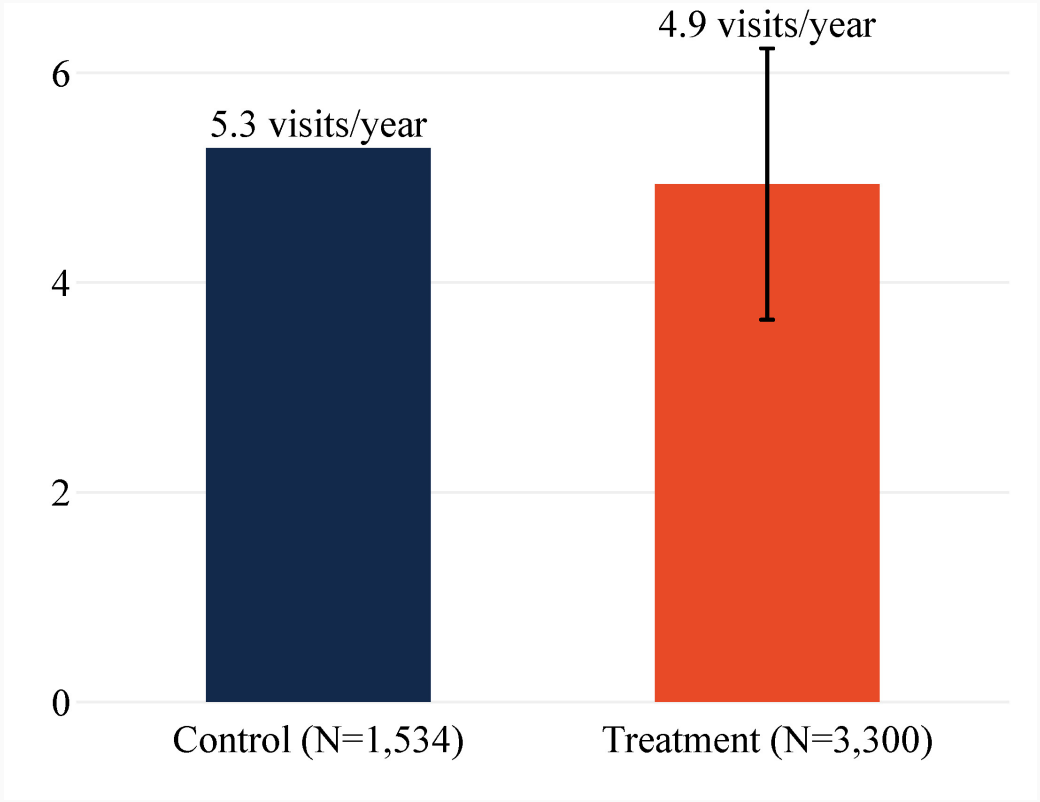


Never had a screening (24-month survey)

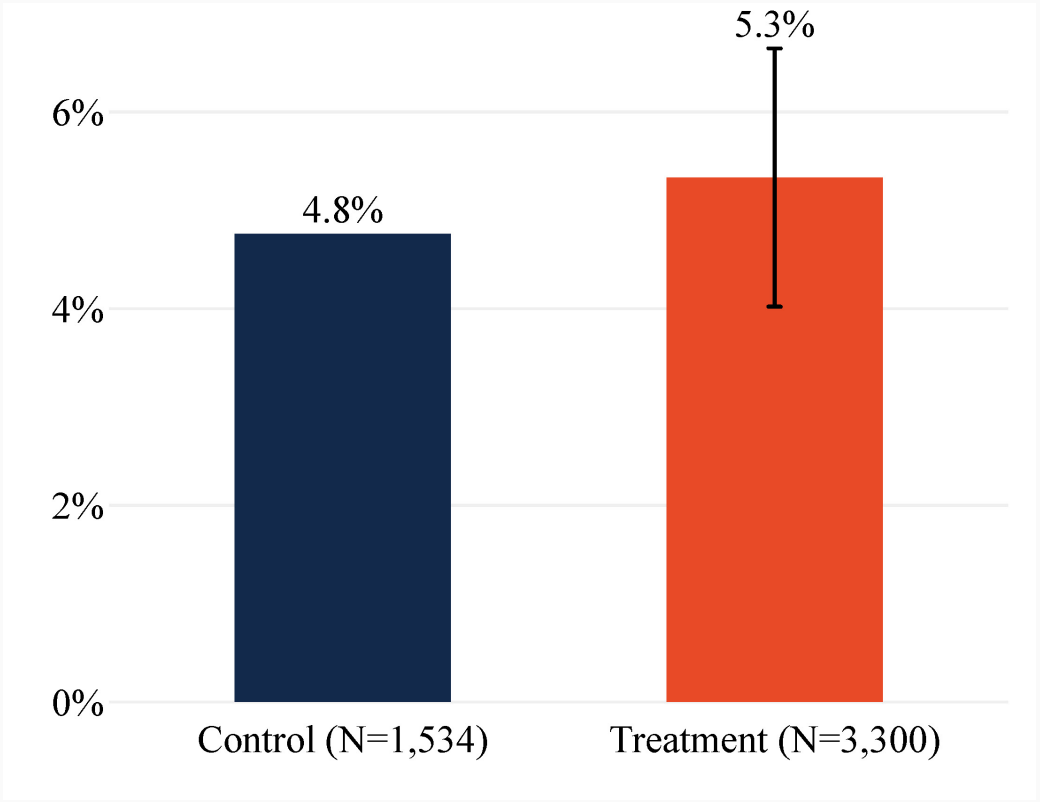
No causal effect on medical spending after 30 months



No causal effect on health behaviors after 30 months

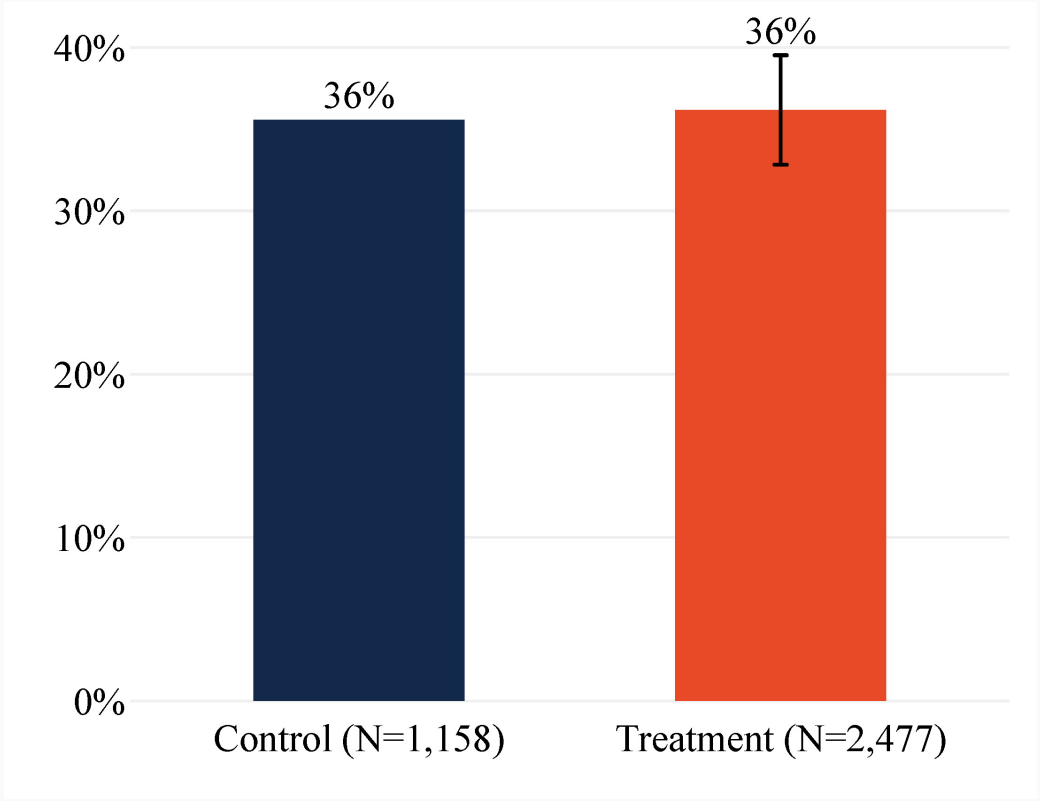


Average annual gym visits

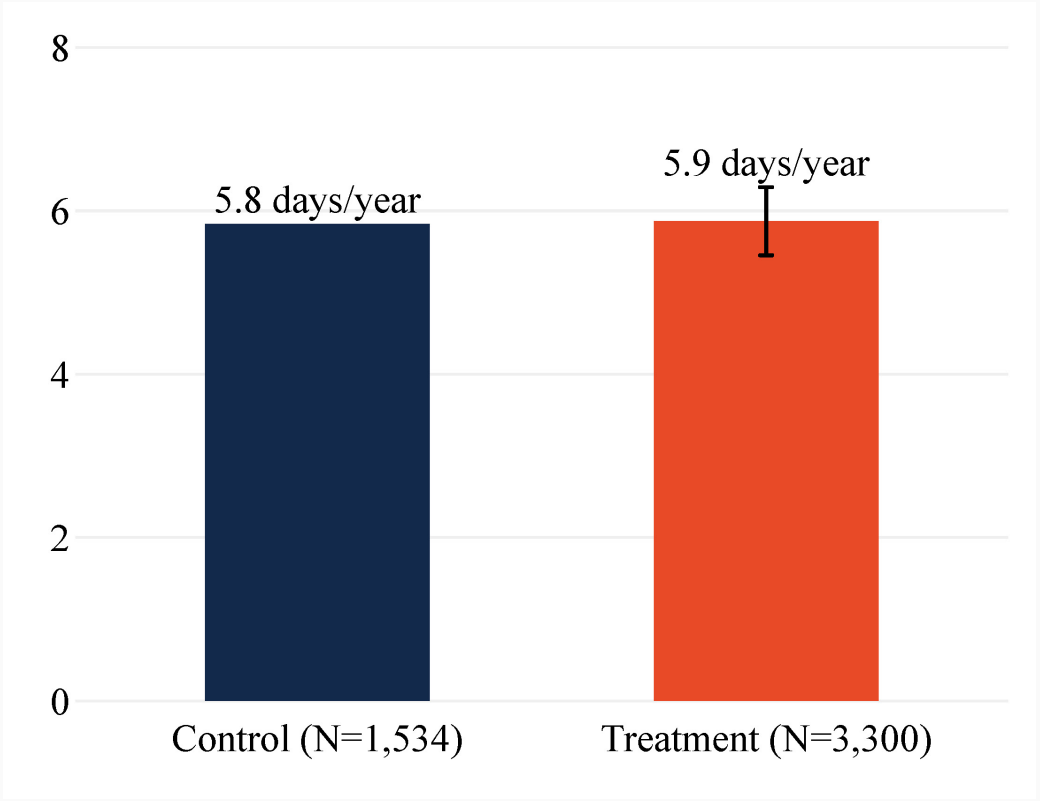


Running event participation

No effect on employee productivity after 30 months

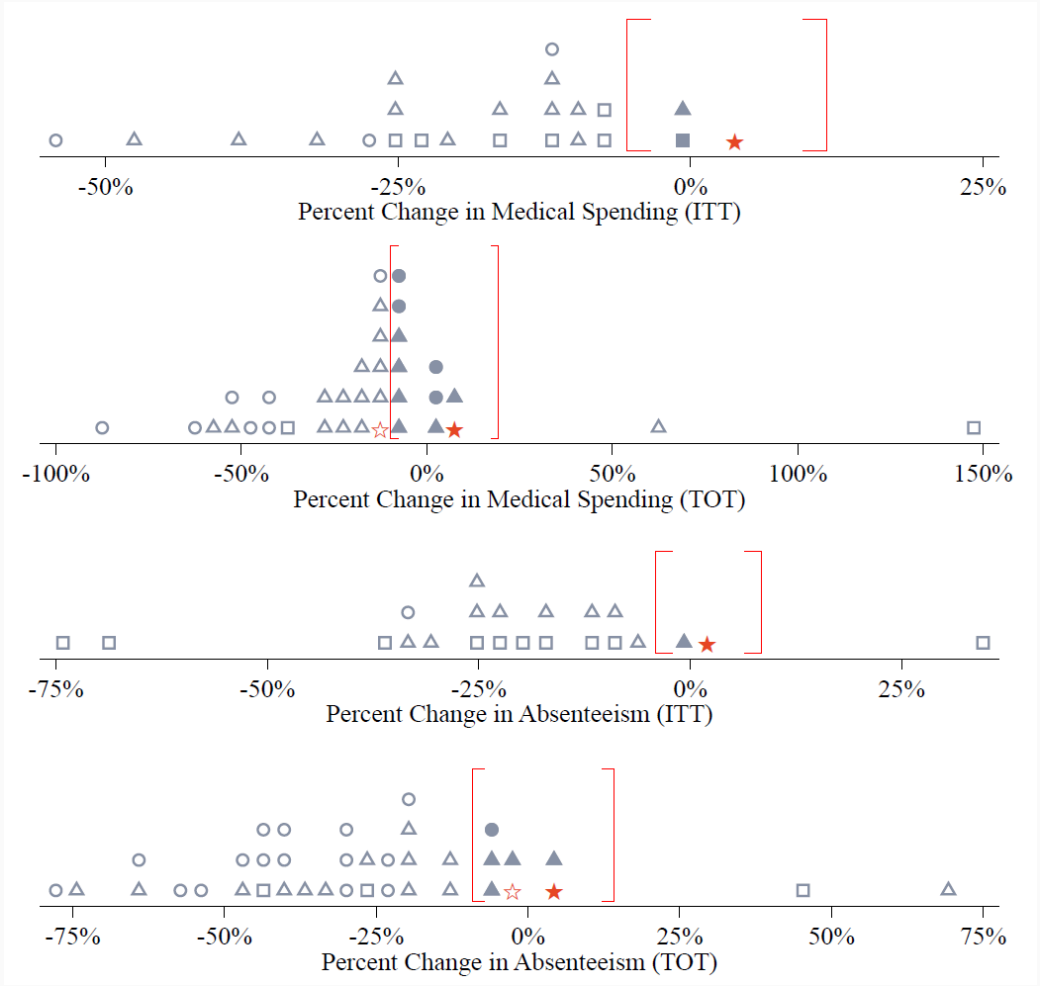


Received job promotion



Sick leave

We rule out 84% of estimates from prior studies



- Is our estimate noisy? Compare to other estimates
- Orange shaded star is our estimate
- Red brackets depict 95% confidence intervals

Why do our results differ from prior studies?

- Randomization was successful
- Our study does not suffer from attrition or reporting bias
- We have good statistical power

- Possible explanations for the discrepancy:
  1. Perhaps our study is not externally valid
  2. Prior observational studies may suffer from selection bias (most are not randomized)

What happens if we estimate a non-randomized model using our data?

Workplace wellness analysis in R

In-class analysis

- IL Workplace Wellness Study datasets and documentation are available online
  - https://github.com/reifjulian/illinois-wellness-data

- Today: work with the IL Marathon dataset
  - IL Marathon: annual running event held in Champaign, IL

- Question: did workplace wellness program increase participation in 2017 IL Marathon?
  - Estimate using randomized experiment
  - Estimate by comparing program participants to non-participants

All variables in the marathon dataset are dummies

Variable name	Definition
treat	1 = treatment group; 0 = control group
hra_c_yr1	1 = completed wellness program; 0 = did not complete
male	1 = male; 0 = female
age50	1 = age $\geq$ 50 (0 otherwise)
age37_49	1 = 37 $\leq$ age $\leq$ 49 (0 otherwise)
white	1 = white; 0 = nonwhite
marathon_2014_2016	1 = marathon participant anytime 2014-2016 (0 otherwise)
marathon_2017	1 = marathon participant in 2017 (0 otherwise)
marathon_2018	1 = marathon participant in 2018 (0 otherwise)

Load and inspect marathon dataset

```r
library("tidyverse")
library("haven")
library("broom")
marathon <- read_dta("https://reifjulian.github.io/illinois-wellness-data/data/stata/marathon.dta")

mean(marathon$marathon_2017) # 6.56% of university employees participated
head(marathon) # `hra_c_yr1` has some missing values. Why?
```

```
# [1] 0.06557716
# # A tibble: 6 × 9
#   treat hra_c_yr1  male age50 age37_49 white marathon_2014_2016 marathon_2017
#   <dbl>     <dbl> <dbl> <dbl>    <dbl> <dbl>               <dbl>         <dbl>
# 1     1         0     0     0        0     1                   0             0
# 2     0        NA     0     0        1     1                   0             0
# 3     1         0     1     1        0     1                   0             0
# 4     0        NA     1     1        0     1                   0             0
# 5     1         1     0     0        0     1                   0             0
# 6     1         1     1     0        0     1                   0             0
# # ℹ 1 more variable: marathon_2018 <dbl>
```

Estimate effect using two different methods

- Outcome: `marathon_2017` (indicator for completing a running event in 2017)
- Estimate effect of wellness program on this outcome using two different methods
- Method 1: compare treatment group to control group
  - Assignment to the treatment group is random, so this estimate is causal
- Method 2: compare wellness program participants to non-participants
  - This method is what most prior researchers used

Balance test for method 1 (randomization): male

```r
# Fraction male, for treatment and control group
marathon %>%  group_by(treat) %>% summarise(mean_male = mean(male))

# Is the difference statistically significant?
cat("\n")
tidy(lm(male ~ treat, data = marathon), conf.int = T, conf.level = 0.95)
```

```
# # A tibble: 2 × 2
#   treat mean_male
#   <dbl>     <dbl>
# 1     0     0.426
# 2     1     0.428
#
# # A tibble: 2 × 7
#   term        estimate std.error statistic   p.value conf.low conf.high
#   <chr>          <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
# 1 (Intercept)  0.426      0.0126    33.7   8.48e-224   0.401     0.450
# 2 treat        0.00189    0.0153     0.124 9.02e-  1  -0.0281    0.0319
```

Causal effect of program on 2017 marathon participation

```
# Is the estimate statistically significant?
tidy(lm(marathon_2017 ~ treat, data = marathon), conf.int = T, conf.level = 0.95)
```

```
# # A tibble: 2 × 7
#   term        estimate std.error statistic  p.value conf.low conf.high
#   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
# 1 (Intercept)  0.0645    0.00632    10.2   3.18e-24   0.0521    0.0769
# 2 treat        0.00152   0.00765     0.199 8.42e- 1  -0.0135    0.0165
```

Balance test for method 2: male

```r
# Fraction male, for people who chose to participate vs those who chose not to
marathon %>% drop_na(hra_c_yr1) %>% group_by(hra_c_yr1) %>% summarise(mean_male = mean(male))

# Is the difference statistically significant? What does that imply?
cat("\n")
tidy(lm(male ~ hra_c_yr1, data = marathon), conf.int = T, conf.level = 0.95)
```

```
# # A tibble: 2 × 2
#   hra_c_yr1 mean_male
#       <dbl>     <dbl>
# 1         0     0.460
# 2         1     0.402
#
# # A tibble: 2 × 7
#   term        estimate std.error statistic   p.value conf.low conf.high
#   <chr>          <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
# 1 (Intercept)   0.460     0.0130     35.5  6.88e-234   0.435     0.485
# 2 hra_c_yr1    -0.0580    0.0173     -3.35 8.24e-  4  -0.0920   -0.0240
```

Method 2: effect of program on marathon participation

```
# Is the effect statistically significant? What does that tell us?
tidy(lm(marathon_2017 ~ hra_c_yr1, data = marathon), conf.int = T, conf.level = 0.95)
```

```
# # A tibble: 2 × 7
#   term        estimate std.error statistic  p.value conf.low conf.high
#   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
# 1 (Intercept)   0.0331   0.00648      5.11 3.49e- 7   0.0204    0.0458
# 2 hra_c_yr1     0.0589   0.00865      6.81 1.15e-11   0.0420    0.0759
```

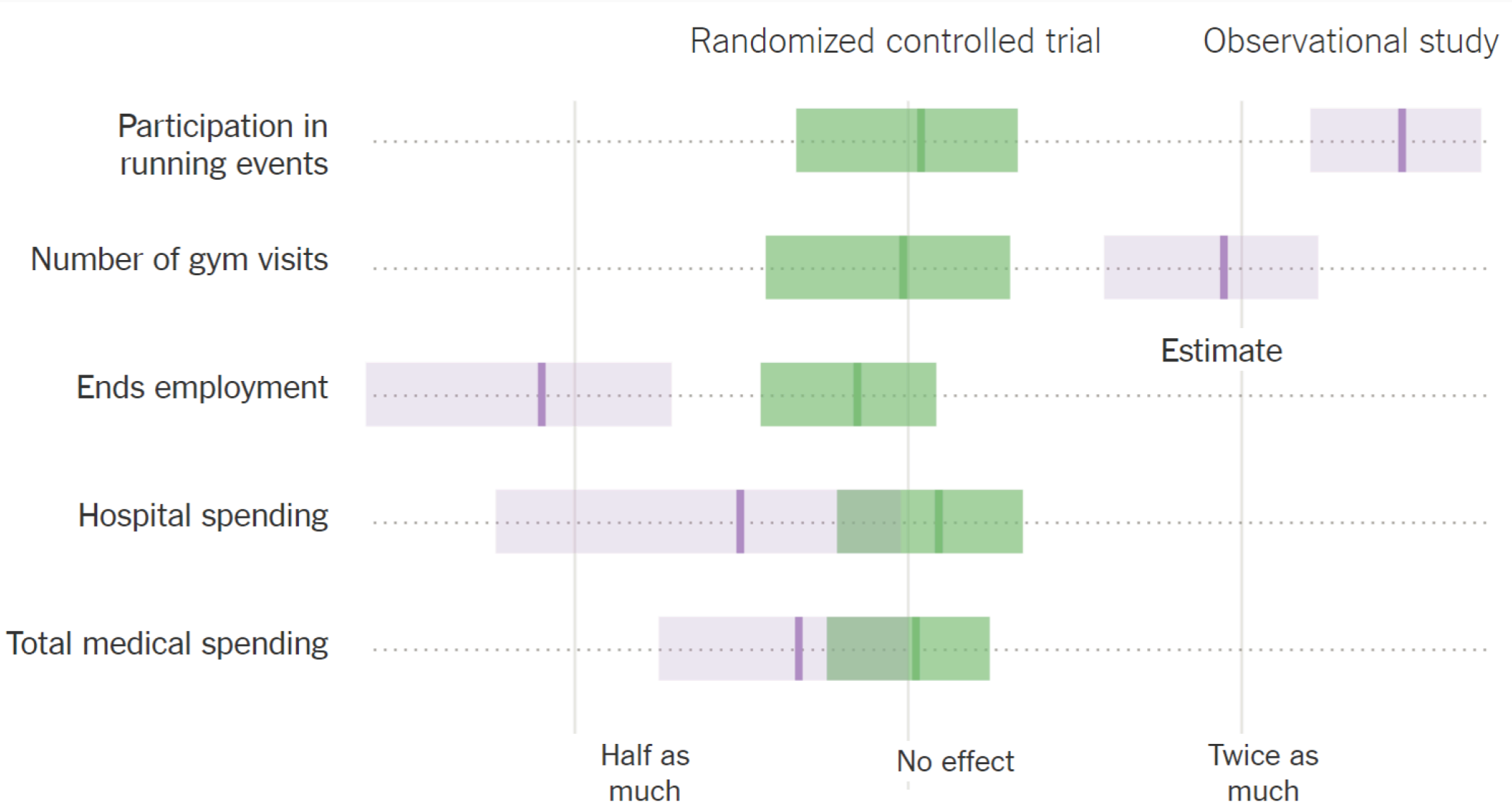
A non-randomized study yields different conclusions!

```r
# Method 1 (randomization): no effect on marathon participation
tidy(lm(marathon_2017 ~ treat, data = marathon), conf.int = T, conf.level = 0.95)


# Method 2 (non-randomization): large effect on marathon participation
cat("\n")
tidy(lm(marathon_2017 ~ hra_c_yr1, data = marathon), conf.int = T, conf.level = 0.95)
```

```
# # A tibble: 2 × 7
#   term        estimate std.error statistic  p.value conf.low conf.high
#   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
# 1 (Intercept)  0.0645    0.00632    10.2   3.18e-24   0.0521    0.0769
# 2 treat        0.00152   0.00765     0.199 8.42e- 1  -0.0135    0.0165
#
# # A tibble: 2 × 7
#   term        estimate std.error statistic  p.value conf.low conf.high
#   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
# 1 (Intercept)   0.0331   0.00648      5.11 3.49e- 7   0.0204    0.0458
# 2 hra_c_yr1     0.0589   0.00865      6.81 1.15e-11   0.0420    0.0759
```

A non-randomized study yields different conclusions!



Lessons from the Illinois Workplace Wellness Study

1. Participants were already healthier and had lower medical spending
    - Wellness programs shift costs onto less healthy employees
    - Wellness programs may be effective way to attract healthy workers

2. No effects on medical spending, health behaviors, or productivity after 30 months
    - We rule out majority of estimates from prior studies

3. Observational health studies likely to suffer selection bias, even with rich controls