Group Members:
Lida Wei
Zien Zhu
Jimmy Chen

Overview of the Case:

- The main focus of the study is to discover the true causal relationship between participating in the Workplace wellness program and Health Condition Improvement. The **Illinois Workplace Wellness Study** uses a **randomized controlled trial (RCT)** design to rigorously assess the impact of a wellness program at the University of Illinois at Urbana-Champaign.
- Workplace Wellness program is designed as a preventive healthcare program aimed at reducing high healthcare spending in the US. Many organizations and companies promote it because many studies show the program is effective at reducing healthcare spending in firms. However, the methods of the studies are questioned by many critics.

Objectives:

1. **Assess the Impact**: To evaluate whether workplace wellness programs can **reduce healthcare spending** and **absenteeism** among employees.
2. **Evaluate Health Outcomes**: To analyze the effect of wellness programs on improving **health behaviors** and the overall well-being of employees.
3. **Address Selection Bias**: To overcome potential **selection bias** using an RCT approach, ensuring that the results accurately reflect the causal impact of the wellness program.

Questions

1. Suppose we want to estimate the effect of workplace wellness programs on employee health care spending. One approach would be to compare health care spending at firms with wellness programs to firms without such programs. What condition would need to be true for this comparison to describe the causal effect of workplace wellness on health care spending? Is this condition likely to be satisfied? Why or why not?

To correctly test the causal effect, it needs to make sure that holding all other variables constant, the only difference should be whether the firms with wellness programs to firms. However, in the approach mentioned in the case, it cannot satisfy the conditions of randomized experiments. It did not hold all the variable constants. The result can be caused by factors like different sectors of industries, different geolocation of companies, different employee demographics, and so on. It will produce a selection bias, which makes it difficult to isolate the wellness program's real influence on healthcare spending.

2. Suppose we want to estimate the effect of workplace wellness among employees in the Illinois study. One approach is to compare outcomes for employees eligible to participate in the wellness program to ineligible employees. What condition would need to be true for this comparison to describe the causal effect of workplace wellness on employee outcomes? Is this condition likely to be satisfied? Why or why not?

To estimate the causal effect of workplace wellness programs on employee outcomes by comparing employees eligible for a wellness program to ineligible employees. The significant condition is there is no selection bias based on unobserved variables.

In other words, eligible and ineligible employees should only differ in wellness program participation status, not in other underlying health or demographic factors that could also influence health outcomes.

However, in the real world, this condition is hard to satisfy because eligibility may often correlate with other factors, such as job type, baseline health status, or access to resources that also impact health. As a result, eligible employees might systematically differ from ineligible ones in ways that bias the estimated effect of the wellness program on health outcomes. Randomization or methods to address selection bias, such as instrumental variables, may help mitigate this issue, but without them, the comparison may not reliably capture the causal effect.

3. What do "treatment" and "control" mean in the context of this study? How does this differ from the definition of a "participant" and "non-participant"? How many employees were in the treatment group, and how many in the control group? Of those in the treatment group, how many participated in the initial (screening) segment of the wellness program in the first year? The answer to this question can be found in the publications of the Illinois Workplace Wellness Study, but use the public-use data (described in Section 3) to form your answers.

Treatment group: employees who were randomly chosen to take part in the workplace wellness program
Control group: employees who were not chosen for the workplace wellness program.

````
# Sum of the TREAT/CONTROL
```{r}
sum(claims$treat == '1') # calculate sum of treat group
sum(claims$treat == '0') # calculate sum of control group
```

[1] 3300
[1] 1534
````

Participant: employees who actively took part in the wellness program after being placed in the treatment group.
Non-participant: employees who did not take part in the program even though they were eligible.

There are 3,300 employees in the treatment group and 1,534 employees in the control group.
What's more, 1900 employees in the treatment group participated in the first year's screening phase of the wellness program.

```{r}
# calculate sum of treatment group participated in the first year's screening phase of the wellness program.
sum(claims$treat == '1'& claims$completed_screening_nomiss_2016 == '1')
```

```
[1] 1900
```

4. Find the "claims" dataset in the public use data repository. For each outcome in the claims dataset, <u>as measured pre-randomization (i.e., prior to August 2016)</u>, report the following in a four-column table (one row per outcome): column (1) Variable description; column (2) Control group mean; column (3) Treatment group mean; and column (4) p-value on the difference. <u>Use linear regression</u> <u>to calculate all these values, and report and describe this equation in your answer below.</u> Do the results in this table support or undermine the claim that treatment group assignment was random?

In conclusion, the linear regression indicated there are insignificant differences in the pre-randomization outcomes between the treatment and control groups. Based on our result, the p-values are quite high (above 0.05), which can support the claims that the treatment group assignment was random.

	Variable.Description	Control.Group.Mean	Treatment.Group.Mean	P.value
1	covg_0715_0716	0.63	0.63	0.8219
2	covg_1015_0716	0.64	0.64	0.7443
3	diabetes_1015_0716	6.39	4.85	0.0707
4	hyperlipidemia_1015_0716	16.55	15.43	0.4148
5	hypertension_1015_0716	14.62	13.23	0.2859
6	pcp_any_office_1015_0716	10.26	10.49	0.8461
7	pcp_any_visits_1015_0716	34.08	31.04	0.0853
8	pcp_total_office_1015_0716	0.40	0.42	0.7654
9	pcp_total_visits_1015_0716	0.84	0.79	0.4689
10	pos_er_critical_1015_0716	0.10	0.10	0.7985
11	pos_hospital_1015_0716	0.05	0.05	0.9125
12	pos_office_outpatient_1015_0716	2.56	2.37	0.0637
13	spendOff_0715_0716	66.71	57.98	0.3771
14	spendRx_0715_0716	103.37	101.27	0.9089
15	spendHosp_0715_0716	283.36	259.33	0.3868
16	spend_0715_0716	505.58	464.81	0.3064
17	nonzero_spend_0715_0716	0.90	0.89	0.2620

5. For each outcome in the "claims" dataset, <u>as measured in the first year following randomization</u>, report the following in a three-column table (one row per outcome): column (1) Variable description, column (2) estimated difference between *treatment and control groups* (no demographic controls) along with standard error in parentheses, column (3) estimated difference between treatment and control groups (with demographic controls) along with standard error in parentheses. Your demographic controls should include indicator variables for sex (male/female), race (white/nonwhite), middle age group (37-49/not 37-49), and oldest age group (50+/not 50+). <u>Use linear regression to calculate all these values, and report and describe this equation in your answer below.</u> What do these estimates indicate about causal treatment effects or selection bias? Should we expect the estimates to differ much between columns (2) and (3)?

The differences between with controls and without control are quite small, which indicates limited selection bias by demographic factors. Therefore, the randomization likely balanced the treatment and control groups well, allowing us to estimate causal treatment effects with confidence. The estimated treatment effects are close to zero and not statistically significant, indicating that the wellness program may not have had a big influence on the spending outcomes.

```
#without demographic control
model_off_1_without <- lm(spendHosp_0816_0717~treat, data = claims)
model_rx_1_without <-lm(spendOff_0816_0717~treat, data = claims)
model_hosp_1_without <-lm(spendRx_0816_0717~treat, data = claims)
model_total_1_without <-lm(spend_0816_0717~treat, data = claims)
model_nonzero_1_without <- lm(nonzero_spend_0816_0717 ~ treat, data = claims)
model_covg_1_without <- lm(covg_0816_0717 ~ treat, data = claims)
model_diabetes_1_without <- lm(diabetes_0816_0717 ~ treat, data = claims)
model_hyperlipidemia_1_without <- lm(hyperlipidemia_0816_0717 ~ treat, data = claims)
model_hypertension_1_without <- lm(hypertension_0816_0717 ~ treat, data = claims)
model_pcp_any_office_1_without <- lm(pcp_any_office_0816_0717 ~ treat, data = claims)
model_pcp_any_visits_1_without <- lm(pcp_any_visits_0816_0717 ~ treat, data = claims)
model_pcp_total_office_1_without <- lm(pcp_total_office_0816_0717 ~ treat, data = claims)
model_pcp_total_visits_1_without <- lm(pcp_total_visits_0816_0717 ~ treat, data = claims)
model_pos_er_critical_1_without <- lm(pos_er_critical_0816_0717 ~ treat, data = claims)
model_pos_hospital_1_without <- lm(pos_hospital_0816_0717 ~ treat, data = claims)
model_pos_office_outpatient_1_without <- lm(pos_office_outpatient_0816_0717 ~ treat, data = claims)
```

```
#with demographic control first year after randomization
model_hosp_1 <- lm(spendHosp_0816_0717~treat+male+white+age37_49+age50, data = claims)
model_off_1<- lm(spendOff_0816_0717~treat+male+white+age37_49+age50, data = claims)
model_rx_1 <- lm(spendRx_0816_0717~treat+male+white+age37_49+age50, data = claims)
model_total_1 <- lm(spend_0816_0717~treat+male+white+age37_49+age50, data = claims)
model_nonzero_1 <- lm(nonzero_spend_0816_0717 ~ treat+male+white+age37_49+age50, data = claims)
model_covg_1 <- lm(covg_0816_0717 ~ treat + male + white + age37_49 + age50, data = claims)
model_diabetes_1 <- lm(diabetes_0816_0717 ~ treat + male + white + age37_49 + age50, data = claims)
model_hyperlipidemia_1 <- lm(hyperlipidemia_0816_0717 ~ treat + male + white + age37_49 + age50, data = claims)
model_hypertension_1 <- lm(hypertension_0816_0717 ~ treat + male + white + age37_49 + age50, data = claims)
model_pcp_any_office_1 <- lm(pcp_any_office_0816_0717 ~ treat + male + white + age37_49 + age50, data = claims)
model_pcp_any_visits_1 <- lm(pcp_any_visits_0816_0717 ~ treat + male + white + age37_49 + age50, data = claims)
model_pcp_total_office_1 <- lm(pcp_total_office_0816_0717 ~ treat + male + white + age37_49 + age50, data = claims)
model_pcp_total_visits_1 <- lm(pcp_total_visits_0816_0717 ~ treat + male + white + age37_49 + age50, data = claims)
model_pos_er_critical_1 <- lm(pos_er_critical_0816_0717 ~ treat + male + white + age37_49 + age50, data = claims)
model_pos_hospital_1 <- lm(pos_hospital_0816_0717 ~ treat + male + white + age37_49 + age50, data = claims)
model_pos_office_outpatient_1 <- lm(pos_office_outpatient_0816_0717 ~ treat + male + white + age37_49 + age50, data = claims)
```

	Variable	estimate_difference_without	estimate_difference_with
1	covg_0816_0717	0.004 (0.015)	0.005 (0.015)
2	diabetes_0816_0717	-1.161 (0.882)	-1.009 (0.868)
3	hyperlipidemia_0816_0717	-0.597 (1.456)	-0.089 (1.38)
4	hypertension_0816_0717	-2.688 (1.366)	-2.271 (1.3)
5	nonzero_spend_0816_0717	-0.008 (0.011)	-0.006 (0.011)
6	pcp_any_office_0816_0717	0.02 (1.197)	0.039 (1.189)
7	pcp_any_visits_0816_0717	-0.852 (1.814)	-0.758 (1.803)
8	pcp_total_office_0816_0717	-0.055 (0.076)	-0.052 (0.075)
9	pcp_total_visits_0816_0717	-0.065 (0.084)	-0.059 (0.083)
10	pos_er_critical_0816_0717	-0.02 (0.018)	-0.02 (0.018)
11	pos_hospital_0816_0717	0 (0.024)	0.002 (0.024)
12	pos_office_outpatient_0816_0717	-0.083 (0.125)	-0.049 (0.119)
13	spend_0816_0717	-31.182 (54.256)	-24.286 (53.768)
14	spendHosp_0816_0717	-10.419 (24.629)	-5.731 (39.965)
15	spendOff_0816_0717	-10.298 (40.259)	-7.609 (8.803)
16	spendRx_0816_0717	-7.844 (8.832)	-9.089 (24.626)

6. For each outcome in the "claims" dataset, as measured in the first year following randomization, report the following in a three-column table (one row per outcome): column (1) Variable description, column (2) estimated difference between *participants and non-participants* (no demographic controls) along with standard error in parentheses, column (3) estimated difference between participants and non-participants (with demographic controls) along with standard error in parentheses. Your demographic controls should include indicator variables for sex (male/female), race (white/nonwhite), middle age group (37-49/not 37-49), and oldest age group (50+/not 50+). Use linear regression to calculate all these values, and report and describe this equation in your answer below. What do these estimates indicate about causal treatment effects or selection bias? Should we expect the estimates to differ much between columns (2) and (3)?

We can see that the differences between people who did not finish hra and the people finishing hra are significant in several variables, which means that although people are all in treatment group, but the people who participated in hra are healthier than the people who did not participate in hra, meaning there is a selection bias. Although, we find that in question 5, the difference is not significant between treatment group and control group, in this question 6, we know that people treated but not finishing hra always spent more money on health than the people treated and finishing hra. The people finishing the hra may spend less even if they did not participate in the health program, as they already are healthier.

All in all, it means that our experiment in the treatment group has selection bias.

```
#calculate estimated difference between Participant and non participant without demographic control using linear regression
model_hosp_part <-tidy(lm(spendHosp_0816_0717~hra_c_yr1, data = claims_clean))
model_off_part <- tidy(lm(spendOff_0816_0717~hra_c_yr1, data = claims_clean))
model_rx_part <- tidy(lm(spendRx_0816_0717~hra_c_yr1, data = claims_clean))
model_total_part <- tidy(lm(spend_0816_0717~hra_c_yr1, data = claims_clean))
model_nonzero_part <- tidy(lm(nonzero_spend_0816_0717 ~ hra_c_yr1, data = claims_clean))
model_covg_part <- tidy(lm(covg_0816_0717 ~ hra_c_yr1, data = claims_clean))
model_diabetes_part <- tidy(lm(diabetes_0816_0717 ~ hra_c_yr1, data = claims_clean))
model_hyperlipidemia_part <- tidy(lm(hyperlipidemia_0816_0717 ~ hra_c_yr1, data = claims_clean))
model_hypertension_part <- tidy(lm(hypertension_0816_0717 ~ hra_c_yr1, data = claims_clean))
model_pcp_any_office_part <- tidy(lm(pcp_any_office_0816_0717 ~ hra_c_yr1, data = claims_clean))
model_pcp_any_visits_part <- tidy(lm(pcp_any_visits_0816_0717 ~ hra_c_yr1, data = claims_clean))
model_pcp_total_office_part <- tidy(lm(pcp_total_office_0816_0717 ~ hra_c_yr1, data = claims_clean))
model_pcp_total_visits_part <- tidy(lm(pcp_total_visits_0816_0717 ~ hra_c_yr1, data = claims_clean))
model_pos_er_critical_part <- tidy(lm(pos_er_critical_0816_0717 ~ hra_c_yr1, data = claims_clean))
model_pos_hospital_part <- tidy(lm(pos_hospital_0816_0717 ~ hra_c_yr1, data = claims_clean))
model_pos_office_outpatient_part <- tidy(lm(pos_office_outpatient_0816_0717 ~ hra_c_yr1, data = claims_clean))
```

```
#Participant and non participant difference with demographic control
model_hosp_part_control <-tidy(lm(spendHosp_0816_0717~hra_c_yr1+male+white+age37_49+age50, data = claims_clean))
model_off_part_control <-tidy(lm(spendOff_0816_0717~hra_c_yr1+male+white+age37_49+age50, data = claims_clean))
model_rx_part_control <- tidy(lm(spendRx_0816_0717~hra_c_yr1+male+white+age37_49+age50, data = claims_clean))
model_total_part_control <- tidy(lm(spend_0816_0717~hra_c_yr1+male+white+age37_49+age50, data = claims_clean))
model_nonzero_part_control <- tidy(lm(nonzero_spend_0816_0717~hra_c_yr1+male+white+age37_49+age50, data = claims_clean))
model_covg_part_control <- tidy(lm(covg_0816_0717 ~ hra_c_yr1 + male + white + age37_49 + age50, data = claims_clean))
model_diabetes_part_control <- tidy(lm(diabetes_0816_0717 ~ hra_c_yr1 + male + white + age37_49 + age50, data = claims_clean))
model_hyperlipidemia_part_control <- tidy(lm(hyperlipidemia_0816_0717 ~ hra_c_yr1 + male + white + age37_49 + age50, data = claims_clean))
model_hypertension_part_control <- tidy(lm(hypertension_0816_0717 ~ hra_c_yr1 + male + white + age37_49 + age50, data = claims_clean))
model_pcp_any_office_part_control <- tidy(lm(pcp_any_office_0816_0717 ~ hra_c_yr1 + male + white + age37_49 + age50, data = claims_clean))
model_pcp_any_visits_part_control <- tidy(lm(pcp_any_visits_0816_0717 ~ hra_c_yr1 + male + white + age37_49 + age50, data = claims_clean))
model_pcp_total_office_part_control <- tidy(lm(pcp_total_office_0816_0717 ~ hra_c_yr1 + male + white + age37_49 + age50, data = claims_clean))
model_pcp_total_visits_part_control <- tidy(lm(pcp_total_visits_0816_0717 ~ hra_c_yr1 + male + white + age37_49 + age50, data = claims_clean))
model_pos_er_critical_part_control <- tidy(lm(pos_er_critical_0816_0717 ~ hra_c_yr1 + male + white + age37_49 + age50, data = claims_clean))
model_pos_hospital_part_control <- tidy(lm(pos_hospital_0816_0717 ~ hra_c_yr1 + male + white + age37_49 + age50, data = claims_clean))
model_pos_office_outpatient_part_control <- tidy(lm(pos_office_outpatient_0816_0717 ~ hra_c_yr1 + male + white + age37_49 + age50, data = claims_clean))
```

	Variable	estimate_difference_without	estimate_difference_with
1	covg_0816_0717	0.139 (0.016)	0.141 (0.016)
2	diabetes_0816_0717	-1.464 (0.987)	-1.056 (0.975)
3	hyperlipidemia_0816_0717	0.345 (1.675)	1.292 (1.593)
4	hypertension_0816_0717	-3.858 (1.54)	-3.258 (1.475)
5	nonzero_spend_0816_0717	0.061 (0.013)	0.052 (0.013)
6	pcp_any_office_0816_0717	0.74 (1.383)	-0.08 (1.377)
7	pcp_any_visits_0816_0717	1.754 (2.091)	0.871 (2.086)
8	pcp_total_office_0816_0717	0.004 (0.085)	-0.031 (0.085)
9	pcp_total_visits_0816_0717	-0.001 (0.096)	-0.042 (0.096)
10	pos_er_critical_0816_0717	-0.018 (0.02)	-0.021 (0.02)
11	pos_hospital_0816_0717	-0.069 (0.029)	-0.068 (0.029)
12	pos_office_outpatient_0816_0717	0.206 (0.142)	0.124 (0.135)
13	spend_0816_0717	-117.163 (59.135)	-135.323 (58.557)
14	spendHosp_0816_0717	-103.053 (45.873)	-113.974 (45.642)
15	spendOff_0816_0717	14.964 (7.374)	11.472 (7.34)
16	spendRx_0816_0717	-26.626 (26.835)	-26.306 (26.943)