Risk outline - Reproducible Version

Survey data overview

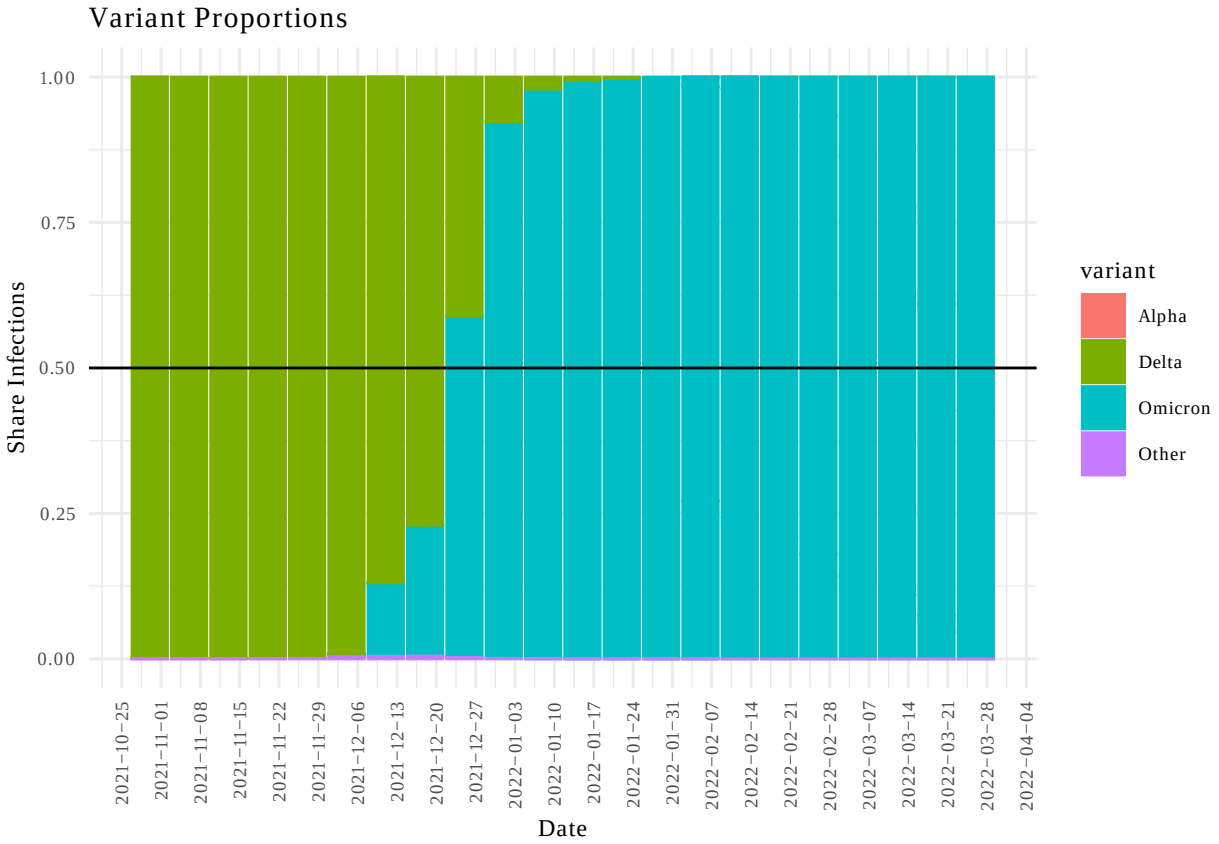
Job risk model

We use Pulse data from phase 3.1, 3.2, and 3.3. The data for these phases pertain to data collected across the following periods:

- Phase 3.1: April 14, 2021 – July 5, 2021
- Phase 3.2: July 21, 2021 – October 11, 2021
- Phase 3.3: December 1, 2021 – December 11, 2021

Pulse surveys prior to phase 3.1 do not contain the question about the job setting of respondents that allows us to categorize employees. Phase 3.3 extends to February 7, 2022 while the new phase¬ 3.4 was conducted from March 2022 to May 2022. However, we exclude these data from the risk model because of the high prevalence of the Omicron variant in these periods. Using CDC data on variant proportions, we see that beginning with the week ending on December 27, 2021, the Omicron variant constituted over half of all COVID infections of the US:

```
read.csv("data_files/covid/covid_variant_proportions.csv") %>%
  filter(modeltype == "smoothed") %>%
  group_by(week_ending,variant) %>%
  arrange(desc(published_date)) %>%
  slice(1:1) %>%
  mutate(week_ending = as.Date(substr(week_ending,1,10),format="%m/%d/%Y"),
         published_date = as.Date(substr(published_date,1,10),format="%m/%d/%Y"),
         variant = case_when(
           variant == "BA.1.1" ~ "Omicron",
           variant == "BA.2" ~ "Omicron",
           variant == "B.1.1.529" ~ "Omicron",
           variant == "BA.2.12.1" ~ "Omicron",
           variant == "B.1.1.7" ~ "Alpha",
           variant == "AY.1" ~ "Delta",
           variant == "AY.2" ~ "Delta",
           variant == "B.1.617.2" ~ "Delta",
           variant == "Other" ~ "Other")) %>%
  ggplot(aes(x=as.Date(week_ending),y=share,fill=variant,col=variant)) +
  geom_col() +
  scale_x_date(date_breaks = "weeks") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  theme(text = element_text(family = "Georgia",size=9)) +
  geom_hline(yintercept=0.5) +
  labs(x = "Date",y = "Share Infections",title = "Variant Proportions")
```



We see that starting the week ending in December 27, 2021, the Omicron variant made up over half of all covid infections in the US. We thus restrict our data to the period ending the week prior.

Creating job categories

We start by assigning a job category to each respondent. We create job categories by combining the type of work the respondent does and their education level. An important point: we only keep respondents who were 18 or older at the time of being surveyed.

1. Work in the last 7 days
    - Variable name: any_work
    - Phase 3.1, 3.2, and 3.3: "In the last 7 days, have you done any work for pay or profit?"
2. Work outside the home
    - Variable name: work_outside_home
    - Phase 3.1: "Since January 1, 2021, have you worked or volunteered outside your home?"
    - Phase 3.2 and 3.3: "In the last 7 days, have you worked or volunteered outside your home?"
3. Setting of work outside home
    - Variable name: setting
    - Phase 3.1: "Since January 1, 2021, which best describes the primary location/setting where you worked or volunteered outside your home?"
    - Phase 3.2 and 3.3: "In the last 7 days, which best describes the primary location/setting where you worked or volunteered outside your home?"
4. Reason not work for pay or profit

- Variable name: reason_not_work
- Phase 3.1, 3.2, and 3.3: "What is your main reason for not working for pay or profit?"

Categorization

- In-person workers
  - Has worked in the last 7 days (any_work = 1)
  - Worked or volunteered outside the home (work_outside_home = 1)
- Remote workers
  - Has worked in the last 7 days (any_work = 1)
  - Has not worked or volunteered outside the home (work_outside_home = 2)
- Unemployed
  - Has not worked in the last 7 days (any_work = 2)
  - Main reason for not working for pay or profit:
    - Was concerned about getting or spreading the coronavirus (reason_not_work = 5)
    - Was laid off or furloughed due to coronavirus pandemic (reason_not_work = 8)
    - Employer closed temporarily due to coronavirus pandemic (reason_not_work = 9)
    - Employer went out of business due to coronavirus pandemic (reason_not_work = 10)
    - I do/did not have transportation to work (reason_not_work = 11)
- NILF
  - Has not worked in the last 7 days (any_work = 2)
  - Main reason for not working for pay or profit:
    - I did not want to be employed at this time (reason_not_work = 1)
    - I am/was sick with coronavirus symptoms or caring for someone who was sick with coronavirus symptoms (reason_not_work = 2)
    - I am/was caring for children not in school or daycare (reason_not_work = 3)
    - I am/was caring for an elderly person (reason_not_work = 4)
    - I am/was sick (not coronavirus related) or disabled (reason_not_work = 6)
    - I am retired (reason_not_work = 7)
    - Other reason, please specify (reason_not_work = 12)

Loading Pulse data, selecting relevant variables, and renaming:

```
clean_vars <- read.csv("data_files/cleanvariables/pulse_cleanvars.csv")
recode_vars <- read.csv("data_files/cleanvariables/pulse_recode.csv")
```

```
phase3.1 <- list.files(path = "data_files/pulse/phase3.1/",
           pattern = "*.csv",
           full.names = T) %>%
    map_df(~read_csv(.)) %>%
  dplyr::select(clean_vars[clean_vars$phase == 3.1,]$pulsename) %>%
  purrr::set_names(clean_vars[clean_vars$phase == 3.1,]$newname)
```

```
phase3.2 <- list.files(path = "data_files/pulse/phase3.2/",
           pattern = "*.csv",
           full.names = T) %>%
    map_df(~read_csv(.)) %>%
  dplyr::select(clean_vars[clean_vars$phase == 3.2,]$pulsename) %>%
  purrr::set_names(clean_vars[clean_vars$phase == 3.2,]$newname)
```

Function to help recode education, setting, race/ethnicity variables:

```
recode <- function(df,x,curr_phase){
  df <- merge(df,recode_vars %>%
          filter(type == x & phase == curr_phase) %>%
          select(-phase) %>%
          dplyr::rename({{x}} := oldvar))
  return(df)
}
```

Clean up data steps:

- Removed respondents born after 2003
- Removed respondents missing responses necessary for categorization
- Categorizing remote workers, nilf, and unemployed
- Recoding variables from numbers to text

```
clean <- function(df, phase){
 nilf <- c(1,2,3,4,6,7,12)
 df <- df %>%
  filter(birth_year <= 2003 & work_outside_home > 0 & setting != -99 & reason_not_work != -99) %>%
  mutate(setting = if_else(setting < 0 & any_work == 1,20,setting),
         setting = if_else(any_work == 2,21,setting),
         setting = if_else(any_work == 2 & reason_not_work %in% nilf,22,setting)) %>%
  filter(setting > 0) %>%
  mutate(race_ethnicity = if_else(hisp_ethnicity == 2,5,race)) %>%
  left_join(unique(get(data("fips_codes")) %>%
              select(state_code,state_name) %>%
              dplyr::rename("state" = "state_code"))) %>%
  select(-state) %>%
  dplyr::rename(state = state_name)

 for(x in unique(recode_vars$type)){
   df <- recode(df,x,phase) %>%
     select(-type,-{{x}}) %>%
     dplyr::rename({{x}} := newvar)
 }
 return(df)
}
```

Obtaining clean data files:

```
phase3.1 <- clean(phase3.1,3.1)
phase3.2 <- clean(phase3.2,3.2)
```

Sanity check Phase 3.2

In survey 3.1, a respondent is remote if:

- Have done any work for pay or profit in last 7 days
- Have not worked or volunteered outside of the home since January 1, 2021

In surveys 3.2 and 3.3, a respondent is remote if:

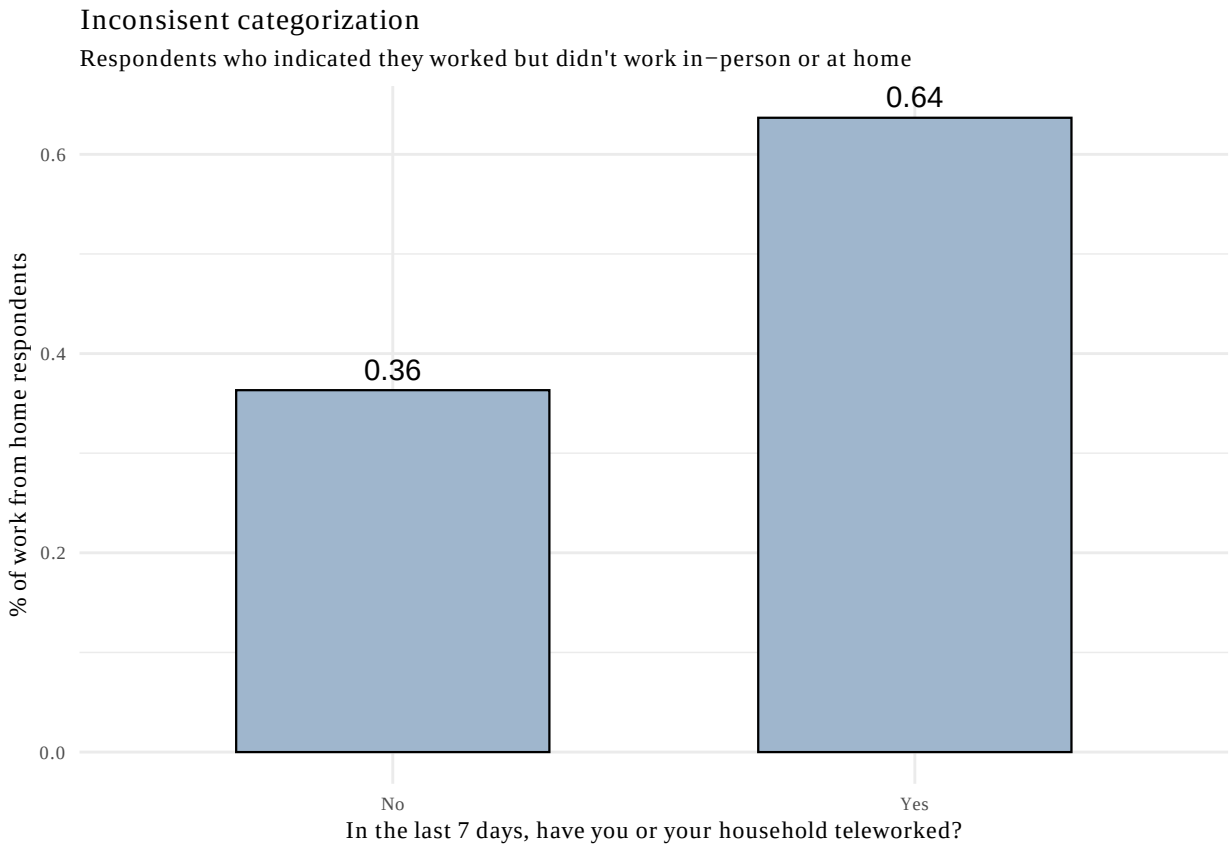
- Have done any work for pay or profit in last 7 days
- Have not worked or volunteered outside of the home in the last 7 days

The phrasing of the questions in the 3.1 phase makes this categorization more defensible since we can more safely assume that someone who has been employed for most of 2021 must be a teleworker if they've done no work outside of the home since January. In 3.2 and 3.3, these employed people who didn't work in-person or remotely might have taken some days off or maybe just happened to not have work that week (client-based workers who didn't have appointments that week, for example).

Doing a sanity check on our 3.2, "In the last 7 days, have you or your household done any of the following... Teleworked or worked from home?" we find that 36% of our "WFH" employees said they didn't work from home.

To handle these "inconsistent" responses, we treat it as missing impute these settings by using a respondent's age, education level, household income, etc. to predict the missing setting.

```
phase3.2 %>%
  filter(setting == "working from home" & teleworked > 0) %>%
  mutate(teleworked = if_else(teleworked == 1,"Yes","No")) %>%
  dplyr::group_by(teleworked) %>%
  dplyr::summarise(count = n()) %>%
  mutate(percent_teleworked = count/sum(count)) %>%
  ggplot(aes(x=teleworked,
             y=percent_teleworked)) +
  theme_minimal() +
  labs(x="In the last 7 days, have you or your household teleworked?",
       y="% of work from home respondents",
       title = "Inconsisent categorization",
       subtitle = "Respondents who indicated they worked but didn't work in-person or at home") +
  geom_col(color = "black",
           fill="slategray3",
           size=0.4,
           width=0.6) +
  geom_text(aes(label=round(percent_teleworked,2)),
            position=position_dodge(width=0.9),
            vjust=-0.5) +
  theme(text = element_text(family = "Georgia",
                            size=9))
```




```
ggsave("figures/inconsisent_responses.png")
```

```
temp <- phase3.2 %>%
  dplyr::mutate(setting = as.character(setting),
                setting = if_else((setting == "working from home" & teleworked == 2),"missing",setting))
  replace_with_na(replace = list(setting = "missing"))
```