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:

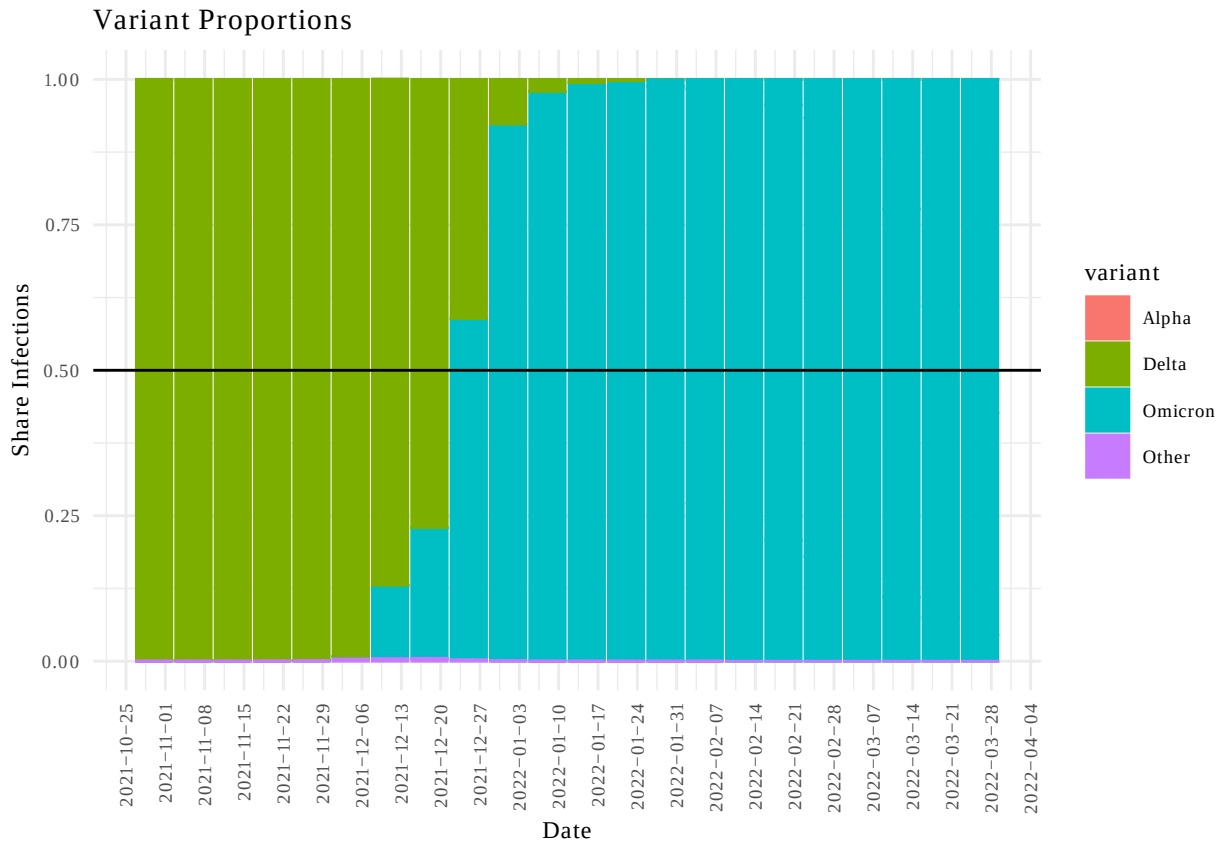
- Phase 3.1: April 14, 2021 – July 5, 2021
- Phase 3.2: July 21, 2021 – October 11, 2021
- Phase 3.3: December 1, 2021 – February 7, 2022
- Phase 3.4: March 2, 2022 - May 9, 2022

Pulse surveys prior to phase 3.1 do not contain the question about the job setting of respondents that allows us to categorize employees.

Omicron concerns + impetus for creating risk of infection by specific severity levels

The period starting from mid-Phase 3.3 and continuing for phase 3.4 contains data collected during a time of high prevalence of the Omicron variant. 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")
```



Because of the highly contagious nature of the variant, we worried readers might think that merely getting infected isn't a big deal given the high share of adults infected at least once. We are also interested in measuring the absolute amount of noxiousness and how it compares to the pre-pandemic period, which means we need a risk measure that is comparable to the types of workplace risk that existed prior to the pandemic. It makes sense to think of an injury ending in death in the same way we think about a covid infection ending in death, but comparing a non-fatal injury such as a cut or laceration with a mild case of covid, or a bad fracture with a case of long covid, is less straightforward. We resolve this apples-to-apples issue by converting the risk of a covid infection to the risk of a mild to moderate infection or a covid hospitalization. We also calculate the risk for the infection to result in a case of long covid. For each type of infection severity, we can give an estimate for the number of days away from work resulting from it by using data on the average recovery time for each type. Finally, we use BLS data collected prior to the pandemic that records the total number of injuries recorded in each industry and the respective recovery times. The bottomline: if a covid infection contracted on the job took 15 days to recover from, we argue the severity of the infection is comparable to a workplace injury that resulted in the worker taking 15 days away from the job.

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, 3.3 and 3.4: "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, 3.3, and 3.4: "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, 3.3, and 3.4: "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, 3.3, and 3.4: "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)

Education Levels We create 4 education categories:

1. Less than high school graduate
2. High school graduate or GED
3. Some college or associate's degree

4. Bachelor's degree or higher

Combining education and setting data

For the categories, "Working from home", "Unemployed", and "NILF", we collapse across all 4 education categories as the type of work and riskiness of it should not vary in these categories. Due to the relatively small size of some of these job categories, we also make the following changes to our job categories:

1. Combine the "less than high school" and "high school" categories for "correctional facilities".
2. Combine the 4 "health care" and 4 "death care" categories into 4 "health and death care" categories.
3. Combine the "less than high school" and "high school" categories for public transit.
4. Combine the "less than high school" and "high school" categories for USPS.

We arrive at a final list of 60 job categories.

Setting	Education
Healthcare and death care	
Social service	
Preschool or daycare	• Less than high school graduate
K-12 school	• High school graduate or GED
Other schools and instructional settings	• Some college or associate's degree
First response	• Bachelor's degree or higher
Food and beverage store	
Agriculture, forest, fishing, or hunting	
Food manufacturing facility	
Non-food manufacturing facility	
Other job deemed "essential" during the COVID-19 pandemic	
None of the above	
Public transit	
Public transit	• High school graduate, GED, or less
United States Postal Service	• Some college or associate's degree
Correctional facility	• Bachelor's degree or higher
Working from home	
Unemployed	• Any education level
NILF	

Figure 1: Job categories

Loading Pulse data, selecting relevant variables, renaming, and cleaning:

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

```
open <- function(phase){
  filepath = paste0("data_files/pulse/phase",as.character(phase),"/")
  df <- list.files(path = filepath,
           pattern = "*.csv",
           full.names = T) %>%
    map_df(~read_csv(.)) %>%
  dplyr::select(clean_vars[clean_vars$phase == phase,]$pulsename) %>%
  purrr::set_names(clean_vars[clean_vars$phase == phase,]$newname)
  return(df)
}
```

```
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)
}
```

Make job categories:

Combining setting information with education level to create job categories.

```
process <- function(phase){
  df <- open(phase)
  df <- clean(df,phase)
  return(df)
}
```

Obtaining clean data files:

```
phase3.1 <- process(3.1)
phase3.2 <- process(3.2)
phase3.3 <- process(3.3)
phase3.4 <- process(3.4)
```

Exporting processed files:

```
write.csv(phase3.1,"data_files/pulse/processed/phase3.1.csv")
write.csv(phase3.2,"data_files/pulse/processed/phase3.2.csv")
write.csv(phase3.3,"data_files/pulse/processed/phase3.3.csv")
write.csv(phase3.4,"data_files/pulse/processed/phase3.4.csv")
```

Sanity check Phase 3.2, 3.3, and 3.4

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, 3.3, and 3.4, 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

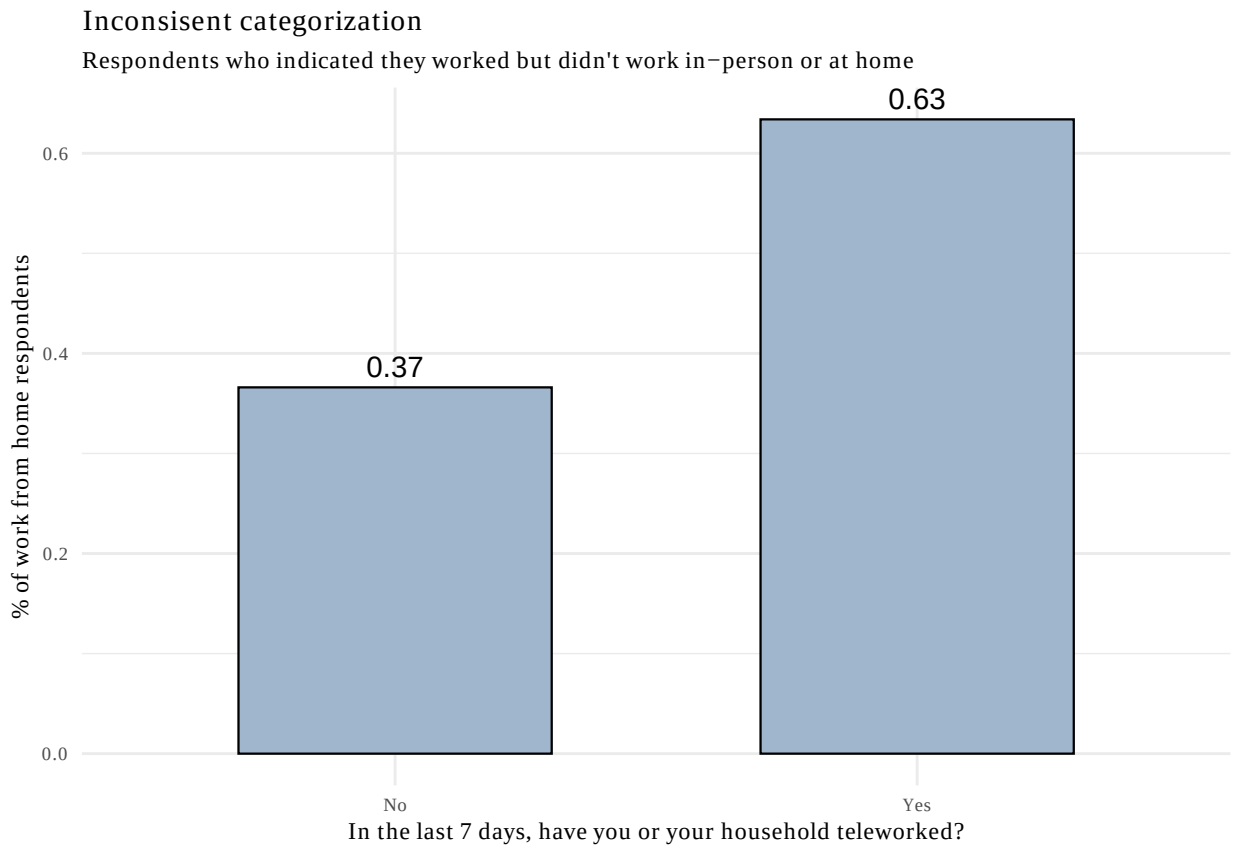
The phrasing of the questions in the 3.1 phase makes this categorization more defensible since we can more safely assume that someone who has been employed for most of 2021 must be a teleworker if they've done no work outside of the home since January. In 3.2 and 3.3, these employed people who didn't 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 the latter phases, "In the last 7 days, have you or your household done any of the following... Teleworked or worked from home?" we find that 37% 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.

```
temp_merge <- rbind(phase3.2,phase3.3,phase3.4)
temp_merge %>%
  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_merge <- temp_merge %>%
  dplyr::mutate(setting = as.character(setting),
                setting = if_else((setting == "working from home" & teleworked == 2),
                                  "missing",
                                  setting)) %>%
  replace_with_na(replace = list(setting = "missing"))
```

```
count(temp_merge$setting)
```

```
##                             x   freq
## 1  agriculture, forestry, etc   5354
## 2       correctional facility   1135
## 3                  death care    558
## 4              first response   4711
## 5     food and beverage store  10863
## 6          food manufacturing   2218
## 7                  healthcare  44393
## 8                k-12 school  25007
## 9      non-food manufacturing   9763
## 10          none of the above 102787
## 11        other essential job  51134
## 12              other schools  15765
## 13       preschool or daycare   3512
## 14             public transit   1838
## 15             social service  14844
## 16                       usps   1004
## 17          working from home 133952
## 18                       <NA>  69881
```