Risk-of-infection outline

# Summary

To obtain an estimate of the number of covid infections happening at work, we take the following high-level steps:

1. Obtain Household Pulse Survey data from phases 3.1, 3.2, 3.3, and 3.4 spanning from April 14, 2021 to May 9, 2022
2. Categorize all Pulse respondents into a setting x education job category
3. Obtain relative covid risk by setting for respondents in each job
   1. Estimate the number of hours the respondent spends in each setting
   2. Estimate how risky those hours are (whether respondent lives alone, whether they practice preventive behavior)
   3. Create 3 subrespondents for each respondent, one for each setting. If respondent is infected, divide infection across subrespondent based on likelihood of contracting in each setting (using hours in setting and risky behavior)
   4. Fit the log-linear model:

*(Job \* Race \* State) + (Job \* Infection) + (State \* Race \* Infection)*

1. Obtain relevant totals:
   1. Total infections (across all settings) by race and state for each period
      1. Combination of CDC and Census data
   2. Total population in each job for each state and racial-ethnic group
      1. Using CPS and ACS data and imputing missing data, smoothing estimates using LOESS, and fitting data to known employment status totals using IPF
   3. Total infections in each setting for all US population
      1. Using CDC study on proportion of infections in each setting and combining it with ATUS data to adjust proportions for each period
2. Use IPF to fit expected values coming out of model to known marginals. This will give us the total number of monthly infections for each job x state x racial-ethnic group.
3. Obtain the risk-of-infection for every respondent and convert this to the risk of mild or moderate infection and severe infection using relevant covariates.

# Survey data overview

Job risk model

We use Pulse data from phase 3.1, 3.2, 3.3, and 3.4. 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 3.4 is the latest data collection period that has been fully released. 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

**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](https://data.cdc.gov/Laboratory-Surveillance/SARS-CoV-2-Variant-Proportions/jr58-6ysp) 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.

Chart, histogram

Description automatically generated

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.

Monthly populations

We use CPS starting March 2020 and ACS data from 2019 to calculate monthly populations for each job by racial-ethnic category.

1. **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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Phase 3.1 | Phase 3.2 | Phase 3.3 |
| Question A – Work in last 7 days | In the last 7 days, have you done any work for pay or profit? | | |
| Question B – Work outside the home | Since January 1, 2021, have you worked or volunteered outside your home? | In the last 7 days, have you worked or volunteered outside your home? | |
| Question C – Setting of work outside home | Since January 1, 2021, which best describes the primary location/setting where you worked or volunteered outside your home? | In the last 7 days, which best describes the primary location/setting where you worked or volunteered outside your home? | |
| Question D – Reason not work for pay or profit | What is your main reason for not working for pay or profit?  1) I did not want to be employed at this time  2)I am/was sick with coronavirus symptoms or caring for someone who was sick with coronavirus symptoms  3) I am/was caring for children not in school or daycare  4) I am/was caring for an elderly person  5) I was concerned about getting or spreading the coronavirus  6) I am/was sick (not coronavirus related) or disabled  7) I am retired  8) I am/was laid off or furloughed due to coronavirus pandemic  9) My employer closed temporarily due to the coronavirus pandemic  10) My employer went out of business due to the coronavirus pandemic  11) I do/did not have transportation to work  12) Other reason, please specify | | |

In-person workers

* Has worked in the last 7 days (Question A = YES)
* Worked or volunteered outside the home (Question B = YES)

Remote workers

* Has worked in the last 7 days (Question A = YES)
* Has not worked or volunteered outside the home (Question B = NO)

Unemployed

* Has not worked in the last 7 days (Question A = No)
* Main reason for not working for pay or profit:
  + Was concerned about getting or spreading the coronavirus (Question D = 5)
  + Was laid off or furloughed due to coronavirus pandemic (Question D = 8)
  + Employer closed temporarily due to coronavirus pandemic (Question D = 9)
  + Employer went out of business due to coronavirus pandemic (Question D = 10)
  + I do/did not have transportation to work (Question D = 11)

NILF

* Has not worked in the last 7 days (Question A = No)
* Main reason for not working for pay or profit:
  + I did not want to be employed at this time (Question D = 1)
  + I am/was sick with coronavirus symptoms or caring for someone who was sick with coronavirus symptoms (Question D = 2)
  + I am/was caring for children not in school or daycare (Question D = 3)
  + I am/was caring for an elderly person (Question D = 4)
  + I am/was sick (not coronavirus related) or disabled (Question D = 6)
  + I am retired (Question D = 7)
  + Other reason, please specify (Question D = 12)

**Discussion categorizing NILF and Unemployed**

We don’t include in NILF:

* I was concerned about getting or spreading the coronavirus
* I do/did not have transportation to work

Because people citing these reasons could have simply quit their risky jobs or jobs that were too far away and are currently trying to find remote or safer employment.

We decide to categorize “Other reason, please specify” as NILF to minimize distance between estimates for the NILF and Unemployed population coming out of the CPS in corresponding months.

**Sanity check Phase 3.2 and 3.3**

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 and 3.3 categorizations using the question, *"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:

Chart, box and whisker chart

Description automatically generated

To handle these “inconsistent” responses, we impute these “Unknown” settings by using a respondent’s age, education level, household income, etc. to predict the missing setting.

Setting x Education Level

We combine data about a respondent’s employment status and work setting with information about their education-level to create job categories. Our reasoning here is that the jobs people are performing in certain settings, e.g., healthcare, are different *types* of jobs and will vary in risk depending on the person’s education.

**Setting**

To find the setting for in-person workers, we use their response to question C (“Which best describes the primary location/setting where you worked or volunteered outside your home?”). Phase 3.2 and 3.3 introduced new job settings not available in Phase 3.1, so we collapse these and recategorize them based on those in 3.1. The following are re-categorized as “Healthcare” to maintain consistency in job settings:

* Hospital
* Nursing and residential healthcare facility
* Pharmacy
* Ambulatory healthcare (e.g. doctor, dentist or mental health specialist office, outpatient facility, medical and diagnostic laboratory, home health care)

**Education**

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

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 to a final list of 60 job categories.

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Phase 3.1** | **Phase 3.2** | **Education Levels** |
| Healthcare and death care | * Healthcare * Death care | * Hospital * Nurshing and residential healthcare facility * Pharmacy * Ambulatory healthcare * Death care | * Less than high school graduate * High school graduate or GED * Some college or associate's degree * Bachelor's degree or higher |
| Social service | * Social service | |
| Preschool or daycare | * Preschool or daycare | |
| K-12 school | * K-12 school | |
| Other schools and instructional settings | * Other schools and instructional settings | |
| First response | * First response | |
| Food and beverage store | * Food and beverage store | |
| Agriculture, forestry, fishing, or hunting | * Agriculture, forestry, fishing, or hunting | |
| Food manufacturing facility | * Food manufacturing facility | |
| Non-food manufacturing facility | * Non-food manufacturing facility | |
| Other job deemed “essential” during the COVID-19 pandemic | * Other job deemed “essential” during the COVID-19 pandemic | |
| None of the above | * None of the above | |
| Public transit | * Public transit | | * High school graduate, GED, or less * Some college or associate's degree * Bachelor's degree or higher |
| United States Postal Service | * United States Postal Service | |
| Correctional facility | * Correctional facility | |
| Working from home | * Has done work for pay or profit in last 7 days * Has **not** worked outside the home since January 1, 2021 | * Has done work for pay or profit in last 7 days * Has not worked outside the home in last 7 days * Indicates to have teleworked in last 7 days | * Any education level |
| Unemployed | * Has not done work for pay or profit in last 7 days because (any of below):   + Concerned about getting or spreading the coronavirus   + Laid off or furloughed due to coronavirus pandemic   + Employer closed temporarily due to coronavirus pandemic   + Employer went out of business due to coronavirus pandemic   + Does/did not have transportation to work | |
| NILF | * Has not done work for pay or profit in last 7 days because (any of below):   + I did not want to be employed at this time   + I am/was sick with coronavirus symptoms or caring for someone who was sick with coronavirus symptoms   + I am/was caring for children not in school or daycare   + I am/was caring for an elderly person   + I am/was sick (not coronavirus related) or disabled   + I am retired   + Other reason, please specify | |

1. **Risk-of-infection model**

Having categorized Pulse respondents into their respective work categories, we now want to create a model to estimate the relative covid risk for different workers. Because we are ultimately interested in understanding risk specific to the workplace, we want to distribute infections between those occurring at work, leisure, or at home. To do this, we take the following steps:

Estimate hours in each setting

Under the assumption that spending more time in a certain setting increases ones probability of contracting covid in it, we first go about estimating the number of hours a respondent spends at home, leisure, and the workplace. To do this, we use pooled state-level ATUS data and determine the average number of hours respondents in each work status (NILF, unemployed, employed remote, employed in-person) spend in the three settings.

Idenitifying risk-takers and risk-averse respondents

Next, because the risk of an hour spend in leisure (eg. shopping mall) depends on whether respondents practice preventive behaviors, we obtain an estimate of whether the respondent is a risk-taker or is risk-averse.

While the Pulse does not contain data on respondents’ preventative behavior, it does contain questions on their beliefs about COVID-19 vaccines which have been shown to be correlated with preventive-behavior habits. For example, a [KFF report](https://www.kff.org/coronavirus-covid-19/poll-finding/kff-covid-19-vaccine-monitor-july-2021/) finds that unvaccinated adults “are more likely to say they are *not worried* they personally will get seriously sick from the coronavirus” and were much less likely than vaccinated adults to report that the Delta variant made them more likely to wear a mask or avoid large gatherings. Thus, we create the following prediction model using relevant variables available in both the YouGov and Pulse surveys:

**Outcome variables**

The YouGov survey contains the following questions about mask-wearing:

1. *Thinking about the last 7 days, how often have you worn a face mask at your place of work?*
2. *Thinking about the last 7 days, how often have you worn a face mask inside a grocery store / super- market?*
3. *Thinking about the last 7 days, how often have you worn a face mask inside a clothing / footwear shop?*
4. *Thinking about the last 7 days, how often have you worn a face mask on public transportation?*

With the possible answers:

* 5 = Always
* 4 = Frequently
* 3 = Sometimes
* 2 = Rarely
* 1 = Not at all

We want to make sure our measure of risky behavior pertains exclusively to the leisure setting, so we exclude question 1 from our measure. However, it is still possible that questions 2-4 are contaminated by work-specific behavior if the respondent works in these settings (grocery store, footwear shop, etc). To get around this issue, we identify workers who are likely to work in these locations based on the job they indicated on the survey and treat their answer to the respective location question as missing, imputing it on the basis of other responses.

The Nature study we use to determine our risk-increasing factor defines the mask wearing group as those who indicated they wore a facemask ‘sometimes’, ‘frequently’, or ‘always’. In line with this study, we say someone is risky at work if their average mask wearing score drops below a 3 (‘sometimes’).

An issue with this method is that some respondents might have a low mask score (“Rarely” or “Not at all”) because they have avoided going to the setting altogether. To address this concern, we also consider the questions:

1. In the last 7 days, have you avoided going to shops
2. In the last 7 days, have you avoided taking public transport

If someone has entirely avoided going to shops (‘Always’), we treat it as if they (‘Always’) wore a mask inside grocery store or clothing/footwear shop.

**Predictors**

1. **Party:** Because state-level covid mandates can influence whether someone wears a mask in leisure settings, and because the party of state-leadership is a strong indicator of mask mandates (e.g., of the eleven states that never issued a statewide masking mandate, all are led by Republican governors) we use a variable to indicate whether the governor in the respondent’s state was a democrat or republican.
2. **Vaccine status and beliefs:** We also use a series of variables relating to whether a person is vaccinated or not, if they want to be vaccinated, and their general beliefs about the vaccine, and government/health authorities:

* **isvax:** Have you had the first or second doses of a Coronavirus (COVID-19) vaccine?
  + 0 = “No, neither”
  + 1 = “Yes, one dose” OR “Yes, two doses”
* **wantsvax:** If a Covid-19 vaccine were made available to me this week, I would definitely get it.
  + 0 = 4, 5 (where 5 = strongly disagree)
  + 1 = 1, 2, 3 (where 1 = strongly agree)
* **gov\_skeptic:** I believe government health authorities in my country will provide me with an effective COVID19 vaccine.
  + 0 = 1, 2, 3 (where 1 = strongly agree)
  + 1 = 4, 5 (where 5 = strongly disagree)
* **test\_concerns:** I have concerns there has not been enough testing of vaccines.
  + 0 = No
  + 1 = Yes
* **mistrustvax:** How much do you trust COVID-19 vaccines?
  + 0 = 1, 2, 3 (where 1 = very much)
  + 1 = 4, 5 (where 5 = not at all)
* **side\_effects:** I have concerns about side effects
  + 0 = No
  + 1 = Yes

If a person is vaccinated they are assigned a 0 for wantsvax. I.e., everyone who “wants to be vaccinated” has not been vaccinated.

If a person is vaccinated or wants to be vaccinated they are assigned a 0 for gov\_skeptic, test\_concerns, mistrustvax, and side\_effects. This is also to be consistent with the Pulse survey which will only ask these questions if person is not vaccinated and does not want a vaccine.

Table

Description automatically generated

Risk-taking multiplier

Having estimated whether someone wears a mask in leisure settings, we want to know how much riskier not wearing a mask is while at leisure. We obtain this risk multiplier using this [Nature study](https://www.nature.com/articles/s41467-021-24115-7#ref-CR25) of 198,077 US participants which finds that self-reported ‘sometimes’, ‘most of the time’, and ‘always’ use of face mask was associated with an average 68% reduced risk of predicted COVID-19 compared to individuals who wore face masks none of the time. Findings are consistent with [this](https://www.cdc.gov/mmwr/volumes/69/wr/mm6923e4.htm?s_cid=mm6923e4_w) CDC study that finds mask wearing reduced risk of infection by 70%. In line with the Nature study, we say that, for risk-takers, a hour spent at leisure is 3.125 (1/0.32) times riskier. As for home risk, if someone lives alone we assign a risk multiplier of 0 because hours spent alone at home are not risky.

Covid contraction

Finally, we create a measure of past covid contraction using the following Pulse survey question:

* *Has a doctor or other health care provider ever told you that you have COVID-19?* 
  + *Yes*
  + *No*
  + *Not sure*

"Yes" responses are marked as having had Covid-19 and given a 1, while "No" and "Not sure" are both marked as never having had Covid and given a 0.

Creating subrespondents

Because we want to divide covid risk between home, leisure, and the workplace, we create three subrespondents for every respondent: a “work respondent,” a “leisure respondent,” and a “home respondent”. Each subrespondent is given a weight of one-third.

We create an infection variable (I) with four categories:

* no infection
* infection at work
* infection at leisure
* infection at home

For a respondent who reports to Pulse that they’ve not been infected, all three subrespondents fall into the “no infection” category in our infection variable (I). For a respondent who reports to Pulse that they have been infected, the work subrespondent will fall into the “infection at work” category, the leisure subrespondent will fall into the “infection at leisure” category, and the home subrespondent will fall into the “infection at home” category. Because we don’t want to distribute an infection equally across each setting, we take the following steps to subdivide the single infection into shares reflecting the likelihood that the infection occurs in each setting:

1. Assign number of hours in each setting
2. Assign risk multiplier for each setting
3. Calculate the “proportion” of the infection occurring in a given setting by multipliying hours in the setting by the riskiness multiplier and dividing by total, risky-adjusted hours.
4. Multiply each proportion by 3 so weighted infections still sum to 1.

For example, for a Pulse respondent who is risky during leisure and lives alone:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Respondent | Weight | Time setting | Risky multiplier | No infection | Infection at work | Infection at leisure | Infection at home |
| WorkResp | 1/3 | 8 | 1 | 0 | 1.17 | 0 | 0 |
| LeisureResp | 1/3 | 4 | 3.125 | 0 | 0 | 1.83 | 0 |
| HomeResp | 1/3 | 12 | 0 | 0 | 0 | 0 | 0 |

Expected values

We can now cross-classify Job (J) x Race (R) x State (S) x Infection (I) and fit the weighted log-linear model:

*(J \* R \* S)* *+ (J \* I) + (R \* S \* I)*

J = Job (60 jobs)

R = Racial-ethic group (5 racial-ethnic groups)

S = State (50 states)

I = Infection (No infection, infection at work, infection at leisure, infection at home)

Racial-ethnic group

To determine a person’s race and ethnicity, we use the following questions:

* *What is your race?*
  + *White, Alone*
  + *Black, Alone*
  + *Asian, Alone*
  + *Any other race alone, or race in combination*
* *Are you of Hispanic, Latino, or Spanish origin?*
  + *No, not of Hispanic, Latino, or Spanish origin*
  + *Yes, of Hispanic, Latino, or Spanish origin*

If a person indicates that they are Hispanic, Latino, or of Spanish origin, they are categorized as Hispanic, regardless of race. This results in 5 race/ethnicity groups:

1. Non-Hispanic white,
2. Non-Hispanic Black,
3. Non-Hispanic Asian,
4. Non-Hispanic other or multiple races
5. Hispanic

**Reasons for incorporating a racial-ethnic group variable**

We add racial-ethnic group as an independent variable in the model rather than as part of the job category (e.g., Black, retail with less than HS) because we don’t think that working in retail with less than a high school degree as a Black person is inherently riskier than the same job category for a white person. The concern instead is that a Black person is more likely to be performing these jobs or living in riskier areas within a state (e.g., working in more crowded retail stores than the average white retail worker). The logic is the same as the one that leads us to include a state variable.

Discussion on potential gender variable

We consider adding a gender variable to our risk-of-contraction model. We could think of doing so in two ways:

* As part of the job category. This is what we did with our educational attainment variable because we believed that working in retail with a bachelor's degree or with less than a high school degree actually represents jobs with different levels of risk.
* As another independent variable in our model. This is what we do with our race/ethnicity, for example.

We thought that the gender effect might be closer to the educational attainment effect: the college-educated health worker who is a woman is far more likely to be a nurse, for example, than a college-educated health worker who is a man. With this in mind, we ran the original regression but this time included gender as part of the job category (e.g., healthcare with a college degree female vs healthcare with a college degree male). Upon seeing these results, we see that in many cases, the differences between the male and female variants of the jobs aren’t all that large, and it’s tough to know what to make of them.

However, we do want to bring in gender via differential death rates. [From the CDC](https://www.cdc.gov/pcd/issues/2020/20_0247.htm#:~:text=According%20to%20the%20largest%20body,fatality%20ratio%20is%20approximately%202.4), "According to the largest body of publicly available sex-disaggregated data from global government sources, although no apparent sex differences exist in the number of confirmed cases, more men than women have died of COVID-19". We thus treat gender as we treat age, and only add it as a factor when we convert risk-of-contraction to risk-of-death.

1. **Calculate absolute risk**

The expected values from the above model refer to infections across the Pulse time period that is covered by our data. We next carry out an IPF in each month in which we ensure that all margins accord with known values:

* Race x state x job population margins 🡪 Total number of people in each of the job categories for every racial-ethnic group in each state in each month
* Race x state infection margins 🡪 Total number of Covid infections in each racial-ethnic group in each state
* Infection x setting margins 🡪 What percent of all covid infections in that month occurred at home, leisure, and work

Obtaining race x state x job population margins

**Crosswalk between CPS Industry and Occupation Codes to Pulse jobs**

We need to estimate the number of people in each of our job categories for each racial-ethnic group for a given time period and state. Because we only have Pulse data relating to job settings from April to December 2021, we instead rely on CPS data for our estimates. CPS data does not have the same job settings that we use to create categories for Pulse. Instead, we have to create a crosswalk between Census industry/occupation codes and Pulse settings. Our crosswalk maps each Census Industry and Occupation combination to a Pulse setting.

A. Essential vs non-essential

Because all Pulse settings are considered essential, we first determine which Census industries are considered essential by using [this](https://www.cdc.gov/niosh/topics/coding/essentialworkers/learnmore.html) list compiled by the National Institute for Occupational Safety and Health (NIOSH) which indentifies essential industries based on [an advisory list](https://www.cisa.gov/sites/default/files/publications/ECIW_4.0_Guidance_on_Essential_Critical_Infrastructure_Workers_Final3_508_0.pdf) published by the Cybersecurity and Infrastructure Security Agency (CISA) in December 2020 and maps them to Census industry codes (CICs). The advisory list “identifies workers who conduct a range of operations and services that are typically essential to continued critical infrastructure viability” including those supporting industries such as medical and healthcare, food and agriculture, transportation and logistics, law enforcement, etc. All industries **not** in this list are mapped to the Pulse setting, “None of the above”.

B. NILF/Unemployed

The CPS contains a monthly labor force recode which defines each respondent as:

* Employed – At work
* Employed – Absent
* Unemployed – On layoff
* Unemployed – Looking
* Not in labor force – Retired
* Not in labor force – Disabled
* Not in labor force – Other

We use this to categorize workers as unemployed or NILF.

C. Working from home

[Dingel and Neiman](https://www.nber.org/papers/w26948) created a list of “teleworkable” occupations using responses from two O\*Net surveys. They make their results available [here](https://github.com/jdingel/DingelNeiman-workathome). We create a crosswalk between this dataset (which uses 2010 SOC codes) and our census occupation codes to categorize CPS respondents as in-person or remote workers.

To summarize:

|  |  |
| --- | --- |
| NILF | * Monthly labor force recode = Not in labor force (retired, disabled, other) |
| Unemployed | * Monthly labor recode = Unemployed (On layoff, looking) |
| Remote | * Monthly labor force recode = Employed (At work, absent) * Census occupation code in Dingal and Neiman’s list of teleworkable occupations |
| In-person essential | * Monthly labor force recode = Employed (At work, absent) * Census industry in CISA’s list of essential industries |
| In-person non-essential | * Monthly labor force recode = Employed (At work, absent) * Census industry in CISA’s list of essential industries |

D. Essential industry to setting

From the list of essential industries, we created a list of possible Pulse settings they could be mapped to. We did this by a combination of using the detailed industry’s higher-level category, as well as [industry descriptions from NAICS](https://www.naics.com/search/). For those industries which could be mapped to more than one setting, we chose the setting that would minimize the overall distance between CPS and Pulse estimates (looped through each possible setting permutation). Similarly, we map all Census occupations to potential Pulse settings (though in many cases the mapping is not possible without industry data). For the most part, we end up using occupation data to move some respondents out of settings we were overestimating for (e.g., first response) and into similar settings we were underestimating for (e.g., correctional facility). As above, we list all possible settings and decide which one to use based on what will minimize the distance between CPS and Pulse (within plausible categorizations based on the occupation). In some cases, an occupation will be mapped to different Pulse settings, depending on the respondent’s industry code. For example, an “education and childcare administrator” will not be mapped to “Other schools” if their industry code indicates they work at a preschool/daycare or K-12 school.

The full crosswalk attached at the end of this document as well as an assessment of our crosswalk’s accuracy done by comparing population estimates from the Pulse and the CPS.

**Impute, smooth, and adjust population estimates**

For each monthly CPS, we have to estimate the population of each job x racial-ethnic group in each state, a total of 15,300 estimates. Given the relatively small sizes of certain jobs, we do not have enough data to arrive at these estimates. Instead, we take a 3-step approach to arrive at our estimates:

1. Multiple imputation

[Kleinke and Reinecke (2013)](https://onlinelibrary.wiley.com/doi/full/10.1111/stan.12009) developed a multiple imputation method for count data (non-negative integer values). We use the associated R package, [countimp](https://www.kkleinke.de/countimp/), to impute missing job x racial-ethnic group counts in the CPS data. Specifically, we fit a Poisson model on observed data from each Census region x job category to estimate missing values. E.g., in the following dataset, the variables state, date, and race/ethnicity would lal be used to predict missing population information:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Job | Region | State | Date | Race/Ethnicity | Population |
| K-12 school | 3 | Texas | January 2021 | White | 100,000 |
| K-12 school | 3 | Louisiana | March 2021 | Black | 60,000 |
| K-12 school | 3 | Florida | August 2020 | Asian | NA |

1. LOESS

Next, we perform LOESS smoothing within each job, racial-ethnic group, and state on the post-imputed population data. To prevent LOESS to smooth estimates into negative values, we perform the local regression on logged population estimates and exponentiate these post-smoothing.

3. IPF

Finally, we ensure that our smoothed estimated don’t violate monthly BLS and CPS known-totals­: state-level unemployment, employment, and NILF population figures, and state-level population estimates for each racial-ethnic group. For each state and month, we cross-classify our LOESS-smoothed estimates for each employment status and racial-ethnic group and use IPF to ensure estimates add up to known-totals:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Employed | Unemployed | NILF | Population |
| Asian | X1,1 | … | … | Total Asian |
| Black | … | X2,2 | … | Total Black |
| Hispanic | … | … | X3,3 | Total Hispanic |
| Other | … | … | … | Total Other |
| White | … | … | … | Total White |
|  | Total Employed | Total Unemployed | Total NILF |  |

Then, we scale specific job populations for each racial-ethnic groups to match our new “employed” estimate.

Obtaining race x state infection margins

We want to get, for each state, the number of COVID cases by racial-ethnic group. To get these COVID cases, we get the covid rates from [this CDC dashboard](https://covid.cdc.gov/covid-data-tracker/#demographicsovertime).

The CDC has covid rates across these racial-ethnic groups:

* White alone, non-Hispanic
* Black alone, non-Hispanic
* Asian or Pacific Islander alone, non-Hispanic
* American Indian or Alaskan Native alone, non-Hispanic
* Hispanic

To convert these covid rates to the actual number of cases, we need the population of each racial-ethnic group in each state. We get this data from [Census estimates](https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-detail.html). The data coming out of the Census give us estimates for the total population across these dimensions:

Origin

* Not Hispanic
* Hispanic

Race

* White Alone
* Black or African American Alone
* American Indian or Alaska Native Alone
* Asian Alone, Native Hawaiian, and Other Pacific Islander Alone
* Two or more races

In the CDC data, all racial groups are non-Hispanic, so we replicate this with the Census, using the non-Hispanic estimates for each racial category (and summing across all Hispanic groups to estimate the total Hispanic population). Since the CDC groups Asian and PI groups, we sum these two racial categories from the Census estimates to create the respective population total. This gives us the total number of cases across each racial-ethnic group.

**Case totals in adult populations**

Having total case counts across each racial-ethnic group, we now must divide the case counts between the population that is under 18 and those 18 and over. This is because our model coming out of the Pulse survey pertains only to adults in the US. Just like with covid rates by racial-ethnic groups, we get covid rates by age group from the same [dashboard](https://covid.cdc.gov/covid-data-tracker/#demographicsovertime), this time selecting the “Age-All Groups” filter. This gives us, for each state, the covid case rate across the following groups:

1. Ages 0-4
2. Ages 5-11
3. Ages 12-15
4. Ages 16-17
5. Ages 18-29
6. Ages 30-39
7. Ages 40-49
8. Ages 50-64
9. Ages 65-74
10. Ages 75+

We use the [same Census data](https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-detail.html) to get population estimates for each age group in each state and convert these rates to absolute numbers. Now, we are able to calculate the total number of Covid cases in adults (all age groups except A, B, C, and D). First, we construct the following table for a given time *t* and state *s* using CDC data on covid cases for all age groups by race/ethnicity:

|  |  |  |
| --- | --- | --- |
|  | Covid | No covid |
| White | X1,1 | X1,2 |
| Black | X2,1 | X2,2 |
| Asian or Pacific Islander | X3,1 | X3,2 |
| American Indian or Alaskan Native | X4,1 | X4,2 |
| Hispanic | X5,1 | X5,2 |

We calculate the column “No covid” by subtracting the number of covid cases from the population totals in the Census data. To get the number of Covid cases in adults, we apply iterative proportional rescaling to adjust the margins to those that are known at time *t* in state *s*:

* the number of whites 18 and older
* the number of Blacks 18 and older
* the number of Asians or Pacific Islanders 18 and older
* the number of American Indian or Alaskan Natives 18 and older
* the number of Hispanics 18 and older
* the number of people 18 and older who had Covid
* the number of people 18 and older who didn’t have Covid

At this point, we have weekly case totals for adults in each racial-ethnic category across each state and spanning from March 7, 2020, to November 27, 2021.

**Correcting racial-ethnic group inconsistencies**

We want to use this data as the margins we will use to fit the table of expected values obtained in our logistic regression. However, the racial-ethnic categories used by the CDC and those recorded in the Pulse survey do not match:

The CDC has covid rates across these racial-ethnic groups:

* White alone, non-Hispanic
* Black alone, non-Hispanic
* Asian or Pacific Islander alone, non-Hispanic
* American Indian or Alaskan Native alone, non-Hispanic
* Hispanic

The Pulse survey has the following categories:

* White alone, non-Hispanic
* Black alone, non-Hispanic
* Asian alone, non-Hispanic
* Any other race alone or in combination
* Hispanic

(Note: Just like with the Census categories, the Pulse has separate race and ethnicity categories, but we use ethnicity information to ensure all people with Hispanic ethnicity are only counted in that group)

Then, to align our categories we use [CDC data](https://covid.cdc.gov/covid-data-tracker/#demographics) on the total number of cases by race/ethnicity at a national level which gives us, for each age group, what percent of all cases in the age group occurred in each of the following racial-ethnic groups:

* 1. Hispanic/Latino
  2. American Indian/Alaska native, non-Hispanic
  3. Asian, non-Hispanic
  4. Black, non-Hispanic
  5. Native Hawaiian / Other Pacific Islander, non-Hispanic
  6. White, non-Hispanic
  7. Multiple/Other, non-Hispanic

We use the same CDC data on the total number of cases in each age-group to convert case percentages to absolute numbers for the data above. We can calculate what percentage of the total cases in adults are attributed to each racial-ethnic group. Merging this data with national-level population data, we can say:

* What percentage of all Asian and Pacific Islander cases happened in Asian populations
* What percentage of the Asian and Pacific Islander population in the US is Asian

For a given state, we calculate what percent of all Asian and PI cases are Asian or PI by:

* % Asian cases (of total Asian + PI cases) = PA,S \* (CA,N / PA,N), where:
  + PA,S = (Asian population in state)/(Asian and PI population in state)
  + CA,N = (Asian cases in US)/(Asian and PI cases in US)
  + PA,N = (Asian population in US)/(Asian and PI population in US)
* % PI cases (of total Asian + PI cases) = PPI,S \* (CPI,N / PPI,N)
  + We adjust each of these so they add up to 1 and we use these to separate our Asian from our PI cases for each state and period.

We repeat the same process to calculate what percent of all cases occurred in multiple/other groups and estimate the number of cases for this group. Finally, we group cases happening in the following into a single category, “Any other race alone or in combination”:

* Native Hawaiian / Other Pacific Islander
* American Indian/Alaska native, non-Hispanic
* Multiple/Other, non-Hispanic

**Weekly to monthly cases**

Finally, because our CPS population estimates have a monthly frequency, we aggregate our weekly covid case estimates by month. To do this, we assume covid cases in a week are distributed evenly across each day (e.g., 35 total cases in a week = 5 daily cases). This allows us to aggregate our data monthly, even if some weeks might fall into two different months.

Obtaining infection x setting margins

Finally, we want to estimate what percent of all cases occurred at home, work, and leisure. We use this [CDC study](https://www.cdc.gov/mmwr/volumes/69/wr/mm6926e3.htm?s_cid=mm6926e3_w) to estimate that, from March to May of 2020, 45% of all infections were attributed to at-home risk, 35% were attributed to workplace risk, and 20% were attributed to leisure-and-errands. We use ATUS data from these months to estimate, on average, how much time people spent in each of these settings.

Distribution of time and covid infections – Nation-wide, March-May

|  |  |  |
| --- | --- | --- |
| **Setting** | **% Day** | **% Covid Infections** |
| At-home | 75% | 44% |
| Workplace | 20% | 46% |
| Leisure-and-errands | 5% | 10% |

The CDC paper is really just giving us the relative per-day risks of work, leisure, and home infection:

* Home: 75% of day, 44% of daily infections (44/75=.59)
* Work: 20% of day, 46% of daily infections (46/20=2.3)
* Leisure: 5% of day, 10% of daily infections (10/5=2)

Multiplying through by the lowest risk setting, our new weights are:

* Home: 1
* Work: 3.9
* Leisure: 3.39

Thus, for each time period we can calculate the average time per day spent in each setting and re-obtain an estimate for the percent of covid infections for the setting.