

TA-1 Scenario 3 "Shopping List"

Since we don't have a standalone mechanism for searching for datasets, we decided to go "old school" and create the list of data we need, distribute work assignments to locate the data, and then use ASKEM tools to ingest the data.

Input data

- **Q1 - Q4. All data should be for these ranges:**
 - June 1 - Sept 30 2021: Training period
 - Oct 1 - Jan 1, 2022: Test period
- **Q5 - Q6. All data should be for these ranges:**
 - June - Dec 2021: Training period
 - Jan 1, 2022 - June 1, 2022: Test period
- **Q1:** SIR model: Calibrate the model parameters using data on cases during the 'training period'.
 - **Dataset #1 needed (done?):** COVID cases at national & state level.
 - Should we also do a literature search for R0?
 - Nelson/Holly have a paper on this
 - **Dataset #2 needed (done?):** COVID recovery at national & state level.
- **Q2:** SIR model with hospitalizations/deaths. ("We would like to update our model to include deaths/and or hospitalizations, in order to incorporate data on those outcomes"). Do not consider vaccination or age stratification
 - **Dataset #3 needed (Brian):** COVID hospitalization at national & state level.
 - **Dataset #4 needed (Chunwei, done):** COVID deaths at national & state level
- **Q3:** Update model to include vaccination status.
 - **Dataset #5: (MikeC, done)** Population vaccination as a function of time
 - MikeC did this, it's done, but Dataset #8 does the same job with age stratification.
 - **Dataset #6/7:** Same as #3/#4 but broken down by vaccination status.
 - *Brian doing dataset #6. Done?*
- **Q4:** Add age stratification:
 - **Dataset #8/9/10:** Same as 5/6/7 but also broken down by age group.
 - *Dataset #8, MikeC, done*
 - *Brian doing dataset #9. Done?*
- **Q5:** Reinfection rate.
 - **Dataset:** Reinfection data? (Chunwei, New York state done) Note shifted time range (Jun - Dec 2021)
- **Q6:** No data needed, run ensemble of models
- **Q7 / Q8:** Literature search for all of the above.

Dataset sources

- CDC Data Portal: <https://data.cdc.gov/>
- Health Data Portal: <https://healthdata.gov/browse>
- Google Health: <https://health.google.com/covid-19/open-data/raw-data>

Datasets Found

Datasets #1 / 2: COVID Cases

Method: Went to CDC data portal, searched for “covid”

Datasets:

- [COVID-19 Case Surveillance Public Use Data](#). Broken down by age. Includes whether patient died or was hospitalized.
- Note: One line per COVID-19 confirmed case (94M lines).
- JHU dataset of COVID-19 - <https://github.com/CSSEGISandData/COVID-19>

Method: Followed Holly’s link from Slack

Datasets

- Google Health data - <https://health.google.com/covid-19/open-data/raw-data>
- Clicked on “Epidemiology”, got the CSV (500 MB!) and filtered out USA rows
- Note: **We do not have recovery** data for the USA in this CSV.

Datasets #3 / 6 / 9: COVID Hospitalization

Method: Went to CDC data portal, searched for “covid hospitals”

Datasets:

- [Case surveillance of COVID-19 hospitalizations](#). Data aggregated for 14 states. Broken down by age and vaccination status. Data is *not* broken down by state.
- Rates are hospitalizations per 100,000 people.
- Data is available for all requested time ranges **except** ages 5 - 11 (which only goes back to December 2021).

Dataset #1 also has information about hospitalizations for individual cases. However, the rates in terms of the population in this dataset is probably more useful.

Method: Followed Holly’s link from Slack

Datasets

- Google Health data - <https://health.google.com/covid-19/open-data/raw-data>
- Clicked on Hospitalizations. Got the CSV and filtered out USA rows
- Some data on ICU is missing from March 2021 (doesn’t matter?)

Datasets #4: COVID deaths

Method: Went to health data portal, searched for “covid death”

Datasets:

- <https://healthdata.gov/dataset/United-States-COVID-19-Cases-and-Deaths-by-State-o/hiyb-zgc2> This dataset contains archived aggregate daily counts of COVID-19 cases and death by state. The dataset covers all the required time ranges.

Method: Followed Holly’s link from Slack

Datasets

- Google Health data - <https://health.google.com/covid-19/open-data/raw-data>
- Death data is present along with cases

Datasets #5

Method: Went to Google, searched for “covid vaccinations”, it took like 20 seconds

Datasets:

- OurWorldInData: <https://ourworldindata.org/covid-vaccinations>
- Population data comes from United Nations World Population Prospects. Income groups come from World Bank. Vaccination stats come from national governments
- Dataset contains all dates and all countries, not just US
- It doesn’t have age stratification

Method: Followed Holly’s link from Slack

Datasets

- Google Health data - <https://health.google.com/covid-19/open-data/raw-data>
- Clicked on Vaccinations. Filtered out USA rows. We have data that we need.

Dataset #8

Method: Went to Google, searched for “us covid vaccinations by age group” and clicked on the first non-advertisement hit. Took less than 15 seconds.

Data:

- <https://data.cdc.gov/Vaccinations/Archive-COVID-19-Vaccination-and-Case-Trends-by-Ag/gxj9-t96f>
- Has great age stratification, is US only.
- Covers December 2020 to October 2022

Method: Followed Holly’s link from Slack

Datasets

- Google Health data - <https://health.google.com/covid-19/open-data/raw-data>
- Clicked on By Age [this is 1G!].

Dataset: Reinfection count

Method: Went to health data portal, searched for “reinfection data”

Dataset:

- <https://healthdata.gov/State/New-York-State-Statewide-COVID-19-Reinfection-Data/sjsp-vrf2>
- This dataset reports the number of reinfections in New York State since January 2021. A reinfection is when a person becomes infected with COVID and later becomes infected again.
- Covers data from Jan 2021 to Jan 2023

Dataset:

- <https://www.cdc.gov/mmwr/volumes/71/wr/mm7104e1.htm>
- Found from <https://www.cdc.gov/coronavirus/2019-ncov/your-health/reinfection.html>; wasn't in xDD.
- Covers New York State & California. Data is aggregated across May - Nov 2021, not broken down.

Other Datasets (perhaps decreasingly useful?):

- <https://www.nature.com/articles/s41591-022-02051-3>: Reinfection data from the VA (national data). Includes hazard rates but couldn't locate numbers behind figures.
- <https://doh.wa.gov/sites/default/files/2022-02/421-024-ReportedReinfections.pdf>: Reinfection data from Washington Department of Health
- <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2794886> from Ireland
- <https://coronavirus.health.ny.gov/covid-19-reinfection-data> has data from NY State.
- <https://www.cdc.gov/mmwr/volumes/70/wr/mm7032e1.htm> Kentucky

Practical considerations for measuring the effective reproductive number, R_t

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325187/>

Notes on what to discuss during meeting

1. Recovery: Do we want to only have a recovery rate or do we need raw data on this? If we have the raw number of cases per day then is that enough?

2. Vaccination: We have a number of vaccinations in data. But we don't know how many vaccinated individuals were infected etc. (but we might if we go through case data). Should we just find parameters for this?
3. Reinfection: Do we just need a rate for this? We have some data from NYC and Washington – can we compute a rate from this and use it?

Data to Handoff

Once you have one of these, upload to GitHub and add to README where you got the information from.

Q1: Simple SIR

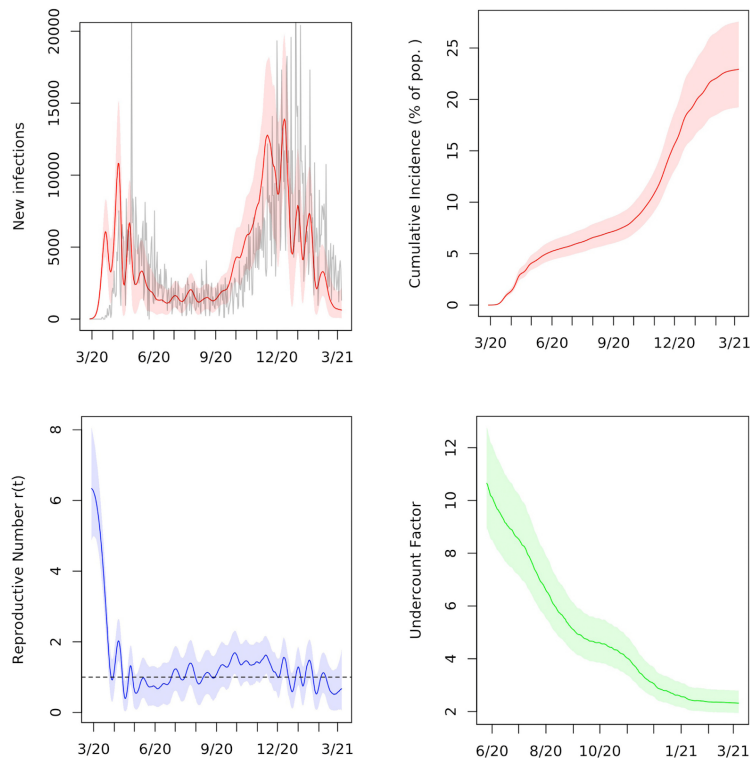
1. Total population of the US (Shiv): From US Census Bureau database. [#26](#)
2. Cases per week since start of pandemic (Shiv): From Google COVID-19 Data [#26](#)
3. **Recovery data (Brian):**
 - a. From “Estimating COVID-19 recovery time in a cohort of Italian healthcare workers who underwent surveillance swab testing” (second hit on TERArium UI for search on “covid recovery time”), 21 days.
 - b. Age stratified medians (in days): Age 25-29: 16.5, 30-39: 23.5, 40-49: 22.5, 50-59: 20.5, 60-66: 25.5
4. **Papers / information justifying conversion** of, e.g., recovery data to a rate in SIR models.

Q2: Add Hospitalization / Deaths

1. Daily time series on number of patients admitted to the hospital all US (Shiv). [#26](#)
2. Daily time series on COVID mortality (Shiv). [#26](#)
3. Monthly time series of hospitalization rate. (Brian, Shiv - cross checked) [#26](#)
 - a. From CDC case data, wrote a quick python script
4. Under-reporting factors, ideally monthly or for periods of interest.
 - a. From <https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html> –
 - i. CDC estimates that from February 2020–September 2021:
 - ii. 1 in 4.0 (95% UI* 3.4 – 4.7) COVID–19 infections were reported.
 - iii. 1 in 3.4 (95% UI* 3.0 – 3.8) COVID–19 symptomatic illnesses were reported.
 - iv. 1 in 1.9 (95% UI* 1.7 – 2.1) COVID–19 hospitalizations were reported.
 - v. 1 in 1.32 (95% UI* 1.29 – 1.34) COVID-19 deaths were reported.
 - b. From <https://www.medrxiv.org/content/10.1101/2020.04.14.20062463v2.full.pdf>
 - i. These prevalence point estimates imply that 54,000 (95CI 25,000 to 91,000 using weighted prevalence; 23,000 with 95CI 14,000-35,000 using unweighted prevalence) people were infected in Santa Clara County by

early April, many more than the approximately 1,000 confirmed cases at the time of the survey (NOTE: April 2020)

- c. Undercount factor is in the bottom right figure (from <https://www.pnas.org/doi/10.1073/pnas.2103272118>)



- d. From <https://journals.asm.org/doi/10.1128/mSystems.00614-20>
 I. Estimating ... suggests that roughly 5% of all samples were positive for SARS-CoV-2 in the 18 to 25 March period, a number much higher than the 0.026% confirmed for the state of Massachusetts (NOTE: Also 2020)
- e. Paper which develops an SEIR model that uses wastewater epidemiology (not sure if there is a specific number we can pull out)
<https://www.medrxiv.org/content/10.1101/2020.11.05.20226738v1.full.pdf>
- f. Another paper that develops a SEIR model that includes waste water data. This one is more recent (from 2022)
<https://www.medrxiv.org/content/10.1101/2022.07.17.22277721v1.full.pdf>
- g. Table 3 in <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7786245/> is great.
<https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2774584> has the link to the PDF

Table 3.

Estimated SARS-CoV-2 Infections, Symptomatic Infections, Hospitalizations, and Deaths by Time Period, 2020

Time period	Reported cases, No.	Infection (symptomatic) underreporting multiplier	Estimated, No.			
			Infections	Symptomatic infections	Hospitalizations	Deaths
January 21-April 30	1 062 446	10.8× (6.5×)	11 474 417	6 905 899	234 801	74 584
May 1-May 31	725 234	4.5× (2.7×)	3 263 553	1 958 132	66 576	21 213
June 1-June 30	837 193	5.4× (3.2×)	4 520 842	2 679 018	91 087	29 385
July 1-July 31	1 917 706	3.9× (2.4×)	7 479 053	4 602 494	156 485	48 614
August 1-November 15	6 303 794	3.2× (1.9×)	20 172 141	11 977 209	407 225	131 119
Total	10 846 373	NA	46 910 006	28 122 752	956 174	304 915

h. Data for 2021 is at

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9716971/pdf/main.pdf>

“From November 4, 2020, to January 27, 2021, the ratio was 2.8 (CI: 2.8–2.9) and then decreased, reaching a low point from April 21 to July 1, 2021 (1.1, CI: 0.6–1.7). These ratios increased to 2.3 (CI: 2.0–2.5) from July 1 to September 20, 2021, held steady from September 20 to December 8, 2021 (2.2, CI: 2.0–2.5), and increased again from December 8, 2021, to February 26, 2022 (3.1, CI: 3.0–3.3).”

the change ratios, ratios estimating the change in seroprevalence compared to the change in reported case prevalence, can be used as a multiplier to enhance the understanding of the infection burden represented by officially reported case rates

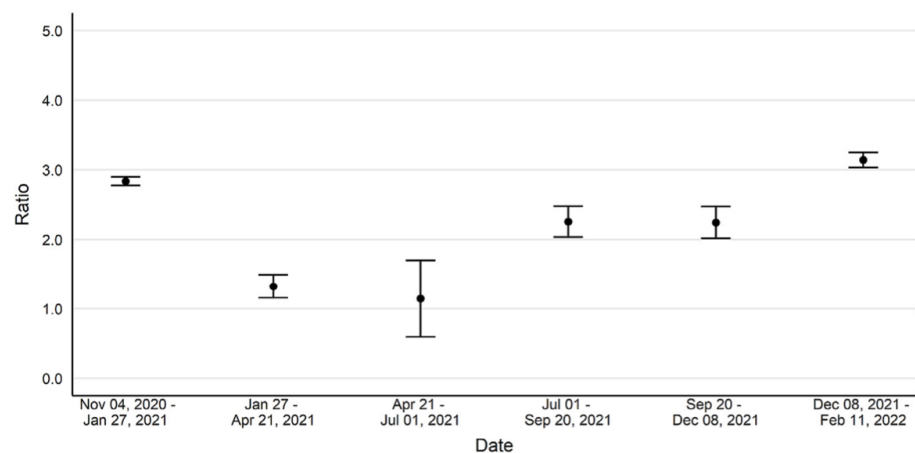


Fig. 3: Estimated change in seroprevalence to change in reported case prevalence ratios for SARS-CoV-2 in the United States, October 25, 2020–February 11, 2022. Data collected as part of a national, repeated, cross-sectional study of convenience samples of specimens of patients who sought routine screening or clinical care. *Footnotes:* Bars represent 95% confidence intervals. Data were collected from the 50 United States

Q3: Vaccination

1. Time series of vaccinations (Shiv) [#26](#)
2. Hospitalization rate difference due to vaccination?
 - a. Searched for “covid vaccine hazard rates” in TERArum. Top hit was CDC MMWR on hospitalization rates. Table 1 has relevant data.
 - b. Using California data from May - Nov 2021: unvaccinated individuals had a 12.7x increased rate of hospitalization compared to vaccinated.
3. **Recovery rate difference due to vaccination?**
 - a. **Nothing apparent in literature search...**
4. Mortality rate difference due to vaccination?
 - a. Went to CDC data tracker, searched for “covid vaccine”, found the dataset “Rates of COVID-19 Cases or Deaths by Age Group and Vaccination Status”
 - b. Wrote a python script to summarize the data by age and month and calculate a hazard rate – the mortality rate for unvaccinated individuals over the mortality rate for vaccinated individuals. [#32](#)

Q4: Age Stratification


- Ensure all datasets in Q1-5 are age-stratified.
 - Cases, Hospitalizations, Deaths are age-stratified [#26](#)
 - Vaccination mortality data is also age-stratified [#32](#)
 - Hospitalization rate is also age-stratified [#32](#)
- Start to stratify parameters as well.

Q5: Reinfection Rate

1. Change in hospitalization for people who are reinfected
2. Median time to reinfection
 - a. Searched for “covid reinfection hazard rates” in TERArum. First hit was the same paper in Q3.
 - b. For California, median time to reinfection is **262 days for vaccinated** and **277 days for unvaccinated**.
 - i. Data was from May - Nov 2021.
 - c. For New York, median time to reinfection is **276 days for vaccinated** and **295 days for unvaccinated**.
 - d. From the VA paper in the literature search (DOI 10.1038/s41591-022-02051-3), the median was 191 days between first and second infection and 158 days between second and third.
 - i. VA data was from 1 June 2020 to 25 June 2022. Not age stratified and the dataset skewed older and male.

All questions

SKEMA Tex Reading extractions that may inform modelers of parameter and value

 scenario3_skema_extractions.xlsx

Data Requirement Notes from TA3 (Chris Rackauckas)

This is a direct dump of the notes written during the discussion. Ping me for any questions, love, or hate mail.

Case data:

- * Expect time series data on I + R
- * Start with an assumption on the recovery
- * Possible additoinal: alternative measure for recovery rate
- * Modeling assumption: use total infections from 2 weeks ago as R_0 , determine I_0 and S_0 from that
- * Need time series for total population of US over time

Deaths and Hospitalizations

- * Daily time series on number of patients admitted to the hospital all US
- * time series for mortality
- * 10 gig file on whether hospitalized or not => percentage for the difference in parameters
 - * Plot the percentage over time by month, see if a constant assumption is okay or not,
 - * If not, need to use the time series
- * Any factor for underreporting estimate? Wastewater time series

Vaccinations

- * Time series of vaccinations
- * Hospitalization rate difference due to vaccination?
- * Recovery rate difference due to vaccination?
- * Mortality rate difference due to vaccination? Hospitalized and not hospitalized

Age-Stratification

- * Previous data that is age stratified is cases, and hospitalizations
- * 10 stratifications, by 10 years each
- * Underreporting over time?
- * Data for assumption on recovery rate with respect to age
- * Aggregated contact matrix for beta over age, from Scenario 1

Reinfection

- * Change in hospitalization for people who are reinfected
- * State of new york, people who reinfected?
- * Median time to reinfection
- * It may require $R \rightarrow S \implies R \rightarrow S_2$
- * Maybe model recovered as vaccinated S ?