

Python Scripts Implementation

1 Overview

This document explains how parameter estimation was done for both the generic SIHRS model and the Carson City case study.

2 Data Extraction

2.1 Hospitalization Data Extraction

Script: `extract_hospitalization_data.py`

This script extracts county-level hospitalization data from the HealthData.gov COVID-19 Hospital Report (2.4 million records, 679 MB). For each county:

- Filters by FIPS code
- Applies date range filter (August 2020 - December 2021)
- Removes invalid data (null/-999999 values)
- Aggregates weekly hospitalization counts
- Saves county-specific CSV files

Output: 2,446 county-level hospitalization CSV files organized by state.

3 Mathematical Formulas and Their Implementation

3.1 Daily Hospitalization Series from Weekly Data

Hospitalization data from the CDC is provided as weekly 7-day averages. To compute daily ratios for p_{IH} estimation, we disaggregate these weekly averages into daily values using a step function method.

3.1.1 Step Function Method

We assign the weekly average value to each day of that week:

$$H_t = H_{\text{week}}, \quad \text{for all } t \text{ in week}$$

where H_{week} is the 7-day average for that week. This creates a step function where the value is constant within each week and jumps at week boundaries.

This method is simple, transparent, and handles genuine zeros and discontinuities in the data (e.g., weeks with zero hospitalizations). For parameter estimation purposes, this step function method produces results that are functionally identical to more complex smooth disaggregation methods, with differences in the final p_{IH} estimate being negligible (less than 0.001%).

3.2 Active Cases Calculation

3.2.1 Mathematical Formula

Active cases are calculated using an *exponential decay method* that aligns with the continuous-time SIHRS model's differential equations. This method is superior to a simple rolling-window approach because it respects the Poisson clock assumption underlying the model.

In the continuous-time SIHRS model, the infectious compartment $I(t)$ evolves according to:

$$\frac{dI}{dt} = C(t) - \gamma I(t),$$

where $C(t)$ is the incidence (new cases per unit time) and γ is the recovery rate. The exact solution is:

$$I(t) = \int_{-\infty}^t C(s) e^{-\gamma(t-s)} ds.$$

For discrete-time data with daily observations, we discretize this integral. Let C_t denote the number of new cases on day t , and let $\gamma = 1/7$ (so $\gamma = 0.1429$, corresponding to a mean recovery time of 7 days). The discrete-time recursive formula is:

$$I_t = C_t + I_{t-1} \cdot e^{-\gamma},$$

where $e^{-\gamma} \approx 0.8669$ is the probability that an individual who was infectious on day $t - 1$ remains infectious on day t (under the exponential distribution assumption).

Given cumulative case counts $\{C_{\text{cum},t}\}$, we first compute daily new cases:

$$C_t = \max(0, C_{\text{cum},t} - C_{\text{cum},t-1}),$$

with $C_1 = C_{\text{cum},1}$. Then we recursively compute active cases:

$$I_1 = C_1, \tag{1}$$

$$I_t = C_t + I_{t-1} \cdot e^{-\gamma}, \quad t = 2, 3, \dots, T. \tag{2}$$

This method ensures that I_t represents the weighted sum of all past new cases, with weights that decay exponentially over time, matching the continuous-time model's dynamics.

3.2.2 Python Implementation

Used in `calculate_pih_all_counties_7day_lag.ma.py` and `calculate_pid_all_counties.py`

```
import numpy as np

# Calculate daily new cases from cumulative cases
county_data['new_cases'] = county_data['cases'].diff().fillna(
    county_data['cases'].iloc[0]
).clip(lower=0)
```

```

# Exponential decay factor: exp(-gamma) where gamma = 1/7
gamma = 1.0 / 7 # recovery_days = 7
decay_factor = np.exp(-gamma) # 0.8669

# Recursive calculation: I[t] = new_cases[t] + I[t-1] * exp(-gamma)
active_cases = []
for i in range(len(county_data)):
    if i == 0:
        active_cases.append(county_data.iloc[i]['new_cases'])
    else:
        current_active = (county_data.iloc[i]['new_cases'] +
                           active_cases[i-1] * decay_factor)
        active_cases.append(current_active)

county_data['active_cases'] = active_cases

```

3.3 p_{IH} : Probability of Hospitalization

3.3.1 Mathematical Formula

We compute p_{IH} using a daily lagged-ratio method with a 7-day moving average. The procedure is:

1. Build a daily hospitalization series H_t by disaggregating weekly 7-day-average hospitalization data using the step function method (see Section 3.1). This method assigns the weekly 7-day average value to all 7 days of that week, creating a step function. We use this method for both the general all-counties analysis and the Carson City case study for consistency.
2. Use NYT case data to construct daily active cases I_t using the exponential decay method (see Section ??) with $\gamma = 1/7 = 0.1429$ (consistent with the model's recovery rate).
3. For each day t , compute the lagged ratio:

$$r_t = \frac{H_t}{I_{t-7}}$$

whenever $I_{t-7} > 0$, where $\tau = 7$ days is the infection-to-hospitalization lag period.

4. Apply a 7-day moving average to smooth the ratio series:

$$\bar{r}_t = \frac{1}{7} \sum_{j=0}^6 r_{t-j}$$

5. Compute the final p_{IH} estimate as the arithmetic mean of all smoothed ratios:

$$p_{IH} = \frac{1}{M} \sum_t \bar{r}_t$$

where M is the number of days with valid smoothed ratios.

3.3.2 Python Implementation

Script: calculate_pih_all_counties_7day_lag.ma.py

The step function method is used for both all-counties analysis and Carson City case study (for consistency):

```
# Build daily H_t from weekly hospitalization data (step function)
daily_hosp = []
for _, row in hosp_weekly.iterrows():
    week_end = row['date']
    value = row['hospitalized'] # weekly 7-day average
    # Assign same value to each of 7 days ending on week_end
    days = pd.date_range(week_end - timedelta(days=6), week_end, freq='D')
    for d in days:
        daily_hosp.append({'date': d, 'H_t': value})

daily_hosp_df = pd.DataFrame(daily_hosp)
```

The remaining steps for p_{IH} calculation (merging with active cases, computing lagged ratios, and applying moving average) are shown below:

```
# Merge with daily active cases I_t
df = pd.merge(daily_hosp_df, active_df.rename(columns={'active_cases': 'I_t'}),
              on='date', how='inner')

# Compute I_{t-7} via 7-day lag
df['I_t_minus_7'] = df['I_t'].shift(7)
df = df[df['I_t_minus_7'] > 0]

# Daily lagged ratio H_t / I_{t-7}
df['ratio_lag7'] = df['H_t'] / df['I_t_minus_7']

# 7-day moving average of the ratio
df['ratio_lag7_ma'] = df['ratio_lag7'].rolling(window=7, min_periods=7).mean()

# Final p_IH as mean of smoothed ratios
df_valid = df[df['ratio_lag7_ma'].notna()]
mean_pih = df_valid['ratio_lag7_ma'].mean()
```

3.4 p_{ID} : Probability of Death

3.4.1 Mathematical Formula

The probability that an infected individual dies from the disease with a 20-day lag:

$$p_{ID} = \frac{1}{n} \sum_{k=1}^n \frac{D_{\text{new}}(t_k)}{I_{\text{active}}(t_k - 20)}$$

where:

- $D_{\text{new}}(t_k)$ = number of new deaths on day t_k
- $I_{\text{active}}(t_k - 20)$ = number of active cases 20 days prior

- n = number of valid data points

3.4.2 Python Implementation

Script: calculate_pid_all_counties.py

```
# Calculate  $P(ID) = \text{daily\_deaths}[T] / \text{active\_cases}[T-20]$ 
lag_days = 20
pid_values = []

for i in range(lag_days, len(county_nyt)):
    date_t = county_nyt.iloc[i]['date']
    daily_deaths_t = county_nyt.iloc[i]['daily_deaths']
    active_cases_t_minus_lag = county_nyt.iloc[i - lag_days]['active_cases']

    if active_cases_t_minus_lag > 0:
        pid = daily_deaths_t / active_cases_t_minus_lag
        pid_values.append(pid)

# Calculate mean  $p_{ID}$ 
mean_pid = np.mean(pid_values)
```

Remark: For p_{ID} , we note that out-of-hospital deaths represent only a small fraction of total deaths (see, e.g., [?, ?]); therefore, we set $p_{ID} = 0$ in the SIHRS model implementation.

3.5 CFR: Case Fatality Rate

3.5.1 Mathematical Formula

The cumulative case fatality rate for the entire study period:

$$CFR = \frac{\text{Total Deaths}}{\text{Total Cases}}$$

where both totals are measured as of December 31, 2021.

3.5.2 Python Implementation

Script: calculate_cfr_all_counties.py

```
# Filter data to study period
period_data = county_nyt[
    (county_nyt['date'] >= '2020-03-01') &
    (county_nyt['date'] <= '2021-12-31')
]

# Get cumulative totals as of end date
total_cases = period_data['cases'].iloc[-1]
total_deaths = period_data['deaths'].iloc[-1]

# Calculate CFR
if total_cases > 0:
    cfr = total_deaths / total_cases
else:
    cfr = 0.0
```

4 National-Scale Analysis

After validation, the scripts were applied to all U.S. counties:

Analysis	Script	Counties	Output
p_{IH} (7-day lag + MA)	<code>calculate_pih_all_counties_7day_lag_ma.py</code>	2,332	Daily ratio calculations
p_{ID}	<code>calculate_pid_all_counties.py</code>	3,218	1,899,366 calculations
CFR	<code>calculate_cfr_all_counties.py</code>	3,140	3,140 values

5 Statistical Summary

For each parameter, the scripts calculate:

- Mean, Median, Standard Deviation
- Percentiles: 25th, 50th (median), 75th, 95th
- Minimum, Maximum
- Individual county values
- For p_{IH} : Interquartile range (IQR) and percentile ranges

Results location: `../03.ProcessedData/`

Note on p_{IH} statistics: Counties with $p_{IH} > 1.0$ (indicating data quality issues or division by near-zero active cases) are excluded from the summary statistics. Out of 2,332 counties processed, 13 outliers were filtered, leaving 2,319 counties for the national summary statistics.

6 Validation: Carson City Case Study

The methodology was first validated on Carson City, NV (FIPS 32510) using the 7-day lagged ratio with 7-day moving average method:

Parameter	Carson City Value	Lag Period
p_{IH} (7-day lag + MA)	0.0920 (9.20%)	7 days
p_{ID}	0.0017 (0.17%)	20 days
CFR	0.0184 (1.84%)	—

The Carson City results match the manual calculations performed in the case study, validating the automated implementation.

For the Carson City case study, we compute p_{IH} using the same methodology but with Carson City-specific data. The empirical distribution yields:

- Mean p_{IH} : 0.1766 (17.66%)
- Median p_{IH} : 0.1454 (14.54%)
- 25th percentile: 0.0891 (8.91%)
- 75th percentile: 0.2240 (22.40%)

- Interquartile range (IQR): [0.0891, 0.2240]

We select $p_{IH} = 0.092$ for the Carson City simulations, which falls at the 25.94th percentile of the distribution, very close to the 25th percentile (0.0891). This choice is justified by the need to account for *ascertainment bias*: underreported COVID-19 infections inflate the empirical ratio H_t/I_{t-7} because the denominator (reported active cases) is smaller than the true number of infections. By choosing a value near the 25th percentile rather than the mean or median, we partially correct for this bias, moving the parameter estimate closer to the true biological probability of hospitalization. This conservative approach also helps prevent the model from over-predicting hospitalization peaks, leading to better calibration with observed data.

7 Key Results (March 2020 - December 2021)

Statistic	p_{IH} (7-day lag + MA)	p_{ID}	CFR
Mean	0.0605	0.0032	0.0177
Median	0.0270	0.0021	0.0168
25th Percentile	0.0037	0.0013	0.0125
75th Percentile	0.0928	0.0036	0.0217
95th Percentile	0.2132	0.0085	0.0316

Note: For p_{IH} , statistics are calculated using the 7-day lagged ratio with 7-day moving average method (see Section 3.3). Outliers with $p_{IH} > 1.0$ (13 counties) were excluded from the summary statistics, resulting in 2,319 counties used for the p_{IH} statistics shown above.

Remark: We suspect that the calculated values of p_{IH} and p_{ID} are slightly higher than the true values, partly because of a low ascertainment rate of COVID-19 infections in the United States (see, e.g., [?, ?, ?]). It also reinforces our assumption of $p_{ID} = 0$ in the SIHRS model implementation.

Additional Remark on p_{IH} :

- Based on age-stratified seroprevalence estimates [?], the infection hospitalization rate ranges from approximately 0.035 (1 in 28.5) for adults 18–49 to 0.082 (1 in 12.2) for adults 50–69. This corresponds to the stated range of 0.04 to 0.08 for p_{IH} in our paper.
- Our chosen value of $p_{IH} = 0.07143$ (1/14) falls within this range. This value is further justified by calculating the infection-weighted average for the entire adult population (ages 18–69) during the Delta variant period using surveillance data [?]:

$$p_{IH} = \frac{249,118 + 321,375}{5,384,265 + 2,991,746} \approx 0.0681 \quad (\text{approximately 1 in 14.6842}).$$

For modeling simplicity and to provide a slightly conservative estimate, we use the nearest simple integer denominator, 1/14, which yields $p_{IH} = 0.07143$.

References

- [1] Kristie EN Clarke. Seroprevalence of infection-induced sars-cov-2 antibodies—united states, september 2021–february 2022. *MMWR. Morbidity and Mortality Weekly Report*, 71, 2022.
- [2] Yangyang Deng, Yun Kim, Anna Bratcher, Jefferson M Jones, Muloongo Simuzingili, Adi V Gundlapalli, Melissa Briggs Hagen, Ronaldo Iachan, and Kristie EN Clarke. Ratio of infections

to covid-19 cases and hospitalizations in the united states based on sars-cov-2 seroprevalence data, september 2021–february 2022. *Open Forum Infectious Diseases*, 12, Oxford University Press US, 2025, p. ofae719.

- [3] Megan O’Driscoll, Gabriel Ribeiro Dos Santos, Lin Wang, Derek AT Cummings, Andrew S Azman, Juliette Paireau, Arnaud Fontanet, Simon Cauchemez, and Henrik Salje. Age-specific mortality and immunity patterns of sars-cov-2. *Nature*, 590(7844):140–145, 2021.
- [4] Elizabeth B Pathak, Rebecca B Garcia, Janelle M Menard, and Jason L Salemi. Out-of-hospital covid-19 deaths: consequences for quality of medical care and accuracy of cause of death coding. 2021, pp. S101–S106.
- [5] Heather Reese, A Danielle Iuliano, Neha N Patel, Shikha Garg, Lindsay Kim, Benjamin J Silk, Aron J Hall, Alicia Fry, and Carrie Reed. Estimated incidence of coronavirus disease 2019 (covid-19) illness and hospitalization—united states, february–september 2020. *Clinical Infectious Diseases*, 72(12):e1010–e1017, 2021.