

Milestone 3

AUTHOR

Group 8: Erin Curlee, Val Stacey, Ana Terzo

PUBLISHED

November 11, 2025

Preparing Data for Analysis

Aligning Column Names

We first ensured that all column names referring to the same variables were consistent across the three databases, using the provided codebook as a reference. Standardizing these names improved both the efficiency of subsequent analyses and the clarity of the data dictionary developed later in this report.

To facilitate merging, we added a county field to the Los Angeles County database so that both datasets share the same set of columns, allowing them to be joined or appended to create a single statewide morbidity dataset for later analyses.

After completing these adjustments — renaming columns in the Los Angeles County database and adding the county field — the remaining tasks involve reconciling the age category and race/ethnicity variable, and standardizing how the timing of infection is identified using MMWR weeks (see **Figure 1** below).

▼ Code

```
##-- California dataset
ca_df <- raw_ca_df %>% select(-dt_diagnosis) %>%
  mutate(county = clean(county) %>%
    str_remove("\s*[Cc]ounty$") %>%
    str_trim())

##-- LA county dataset
la_ctny_df <- raw_la_ctny_df %>%
  rename("age_cat" = "age_category",
        "race_ethnicity" = "race_eth",
        "new_infections" = "dx_new",
        "cumulative_infected" = "infected_cumulative",
        "new_unrecovered" = "unrecovered_new",
        "cumulative_unrecovered" =
        "unrecovered_cumulative",
        "new_severe" = "severe_new",
        "cumulative_severe" = "severe_cumulative") %>%
  mutate(county = "Los Angeles") %>%
  relocate(county, .before = everything())

##-- Population dataset
pop_df <- raw_pop_df %>%
  rename("race_ethnicity" = "race7") %>%
  mutate(county = clean(county))
```

Figure 1. Reconciling Column Names

LA County Dataset			California Dataset	
Original Column Name	New Column Name	Class	Column Name	Class
	county	character	county	character
age_category	age_cat	character	age_cat	character
sex	sex	character	sex	character
race_eth	race_ethnicity	character	race_ethnicity	integer
dt_dx	dt_dx	character	time_int	integer
dx_new	new_infections	integer	new_infections	integer
infected_cumulative	cumulative_infected	integer	cumulative_infected	integer
unrecovered_new	new_unrecovered	integer	new_unrecovered	integer
severe_new	new_severe	integer	new_severe	integer
severe_cumulative	cumulative_severe	integer	cumulative_severe	integer

Restructuring Variables

Age Groups

The population database contains counts for six age groups, whereas the California and LA County databases use four. Because the groupings align — “0–4,” “5–11,” and “12–17” in the population database correspond to the “0–17” group in the others — we summarized the population counts to match the four age-group format.

In the original population database, counts are reported for each single year of age within every demographic subgroup (e.g., county, sex, race/ethnicity). To obtain total estimates per age group and demographic category, the single-year counts were aggregated (summed). **Table 1** below presents a subset of the original and restructured population data to illustrate the resulting summarized age groupings.

▼ Code

```
pop_df <- pop_df %>%
  group_by(county, health_officer_region,
           sex, race_ethnicity, age_cat) %>%
  summarise(pop = sum(pop), .groups = "drop")

pop_df_recat <- pop_df %>%
  mutate(new_age_group =
    if_else(age_cat %in%
      c("0-4", "5-11", "12-17"),
      "0-17", age_cat)) %>%
  group_by(county, health_officer_region, sex,
           race_ethnicity, new_age_group) %>%
  summarise(pop = sum(pop), .groups = "drop") %>%
  rename("age_cat" = "new_age_group")
```

Table 1: Reconciling Age Group Categories and Population Counts

Original Age Groups:

County	Sex	Race / Ethnicity	Age Group	Population
Alameda	FEMALE	Hispanic	0-4	11587
Alameda	FEMALE	Hispanic	12-17	17098
Alameda	FEMALE	Hispanic	5-11	17761
Total	-	-	-	46,446

New, Aggregate Age Group:

County	Sex	Race / Ethnicity	Age Group	Population
Alameda	FEMALE	Hispanic	0-17	46,446

Infection Dates

To support date-based analyses, infection records will be aggregated by **MMWR week and year**. For both the Los Angeles County and California datasets, we will generate two new columns: **mmwr_year** and **mmwr_week**, then remove the original date-based fields. We will also add columns **start_date** and **end_date** to serve as reference points should we need them later.

In the California dataset, the field **time_int** encodes the year and MMWR week as a six-digit integer (YYYYWW). To create the new fields, we extract the first four digits as **mmwr_year** and the last two digits as **mmwr_week**, then drop the original **time_int** column.

In the Los Angeles County dataset, the codebook identifies a field **dt_report** as the last day of the MMWR week. However, this field contained only missing values, so it was removed. Instead, we convert the infection date field, **dt_dx** to a proper date format, and then use the [MMWRweek](#) package to derive the **mmwr_year** and **mmwr_week**.

To streamline this process, we created two helper functions:

- `add_mmwr_week_columns()` : takes date column and adds two fields: **mmwr_year** and **mmwr_week**
- `add_start_end_dates()` : uses those values to generate corresponding MMWR week start and end dates

▼ Code

```
la_ctny_df <- la_ctny_df %>%
##--restructure to proper date format
mutate(DATE_FIX =
  as.Date(parse_date_time(dt_dx, "%d%b%Y"),
         format = "%Y-%m-%d")) %>%
##--use date to create new MMWR fields
add_mmwr_week_columns(date_col = "DATE_FIX") %>%
add_start_end_dates() %>%
select(-c(DATE_FIX, dt_dx)) %>%
relocate(mmwr_year, mmwr_week, start_date,
         end_date, .before = everything()) %>%
relocate(county, .before = age_cat)

ca_df <- ca_df %>%
##--pull MMWR week and year from time_int field
mutate(
  mmwr_year = factor(time_int %% 100),
  mmwr_week = factor(time_int %/% 100)) %>%
add_start_end_dates() %>%
select(-time_int) %>%
relocate(mmwr_year, mmwr_week, start_date,
         end_date, .before = everything())
```

The dataframes now have a structure that looks like this:

mmwr_year	mmwr_week	start_date	end_date	county	age_cat	new_infections
2023	22	2023-05-28	2023-06-03	Los Angeles	0-17	15
2023	23	2023-06-04	2023-06-10	Los Angeles	0-17	17
2023	24	2023-06-11	2023-06-17	Los Angeles	0-17	23

Race and Ethnicity:

Each of the three datasets defines Race / Ethnicity differently. The California dataset uses numeric codes, the Los Angeles County dataset uses full text labels, and the population dataset uses abbreviated text.

To resolve this, we created a crosswalk file (**race_ethnicity_map.csv**) that aligns the three formats. By joining this crosswalk to each dataset, we ensure that all three contain a consistent set of race and ethnicity variables: each with the numeric code, the abbreviated text, and the full text label.

▼ Code

```
ca_df <- ca_df %>%
  rename("race_coded" = "race_ethnicity") %>%
  mutate(race_coded = as.character(race_coded)) %>%
  left_join(race_ethnicity_map, by = "race_coded")%>%
  relocate(race_coded, race_short, race_long, .after = sex)

la_ctny_df <- la_ctny_df %>%
  mutate(race_long = clean(race_ethnicity)) %>%
  select(-race_ethnicity) %>%
  left_join(race_ethnicity_map, by = "race_long") %>%
  relocate(race_coded, race_short, race_long, .after = sex)
```

Race and Ethnicity Crosswalk Table:

race_coded	race_long	race_short
1	White, Non-Hispanic	WhiteTE NH
2	Black, Non-Hispanic	Black NH
3	American Indian or Alaska Native, Non-Hispanic	AIAN NH
4	Asian, Non-Hispanic	Asian NH
5	Native Hawaiian or Pacific Islander, Non-Hispanic	NHPI NH
6	Multiracial (two or more of above races), Non-Hispanic	MR NH
7	Hispanic (any race)	Hispanic
9	Unknown	Unknown

Population Counts by Demographic

To calculate infection rates by demographic groups (such as county, health officer region, sex, or race/ethnicity), we first summarize total population counts for each demographic category within the population dataframe. The population dataset is then joined to both the master database (which merges all three datasets) and the individual California and Los Angeles datasets.

By creating and maintaining these summarized population counts, we avoid having to recalculate them each time we focus on a different demographic group. For example, once this population dataset is joined to the California data, we can easily calculate population-adjusted infection rates that allow valid comparisons across counties with differing population sizes.

▼ Code

```
pop_df <- pop_df_recat %>%
  ##-- join population database to the race/ethnicity map
  mutate(race_short = clean(race_ethnicity)) %>%
  select(-race_ethnicity) %>%
  left_join(race_ethnicity_map, by = "race_short") %>%
  relocate(race_coded, race_short, race_long, .after = sex) %>%
  ##-- calculate population totals by demographic
  group_by(county, health_officer_region) %>%
  mutate(total_ctny_pop = sum(pop)) %>% ungroup() %>%
  group_by(county, health_officer_region, race_coded, race_short, race_long) %>%
  mutate(total_race_pop = sum(pop)) %>% ungroup() %>%
  group_by(county, health_officer_region, sex) %>%
  mutate(total_sex_pop = sum(pop)) %>% ungroup() %>%
  group_by(health_officer_region) %>%
  mutate(total_HOR_pop = sum(pop)) %>% ungroup()
```

Table 2: Calculating Infection Rate by 100,000 Population Example

County	Total Population	Total Infected	Proportion Infected	Infection Rate per 100K
Alameda	1656037	12427	0.0075041	75.0
Alpine	1165	25	0.0214592	214.6
Amador	40122	767	0.0191167	191.2

Final Combined Database

After reconciling all column names across the 3 datasets, standardizing age group, race/ethnicity, and infection date variables, we merge them together to generate a final, complete database and begin our descriptive explorations and analyses.

▼ Code

```
combined_df <- rbind(ca_df, la_cnty_df) %>%  
  relocate(health_officer_region, .after = county) %>%  
  relocate(pop, .after = race_long) %>%  
  mutate(age_cat = factor(age_cat, levels = c("0-17", "18-49", "50-64", "65+")))
```

▼ Code

```
# final cleaned dataframes stored in "_data/cleaned_data/" directory  
  
##--write.csv(combined_df, file = here("_data/cleaned_data/combined_df.csv"), row.names = FALSE)  
##--write.csv(ca_df, file = here("_data/cleaned_data/cleaned_ca_df.csv"), row.names = FALSE)  
##--write.csv(la_cnty_df, file = here("_data/cleaned_data/cleaned_la_cnty_df.csv"), row.names =  
#               FALSE)  
##--write.csv(pop_df, file = here("_data/cleaned_data/cleaned_pop_df.csv"), row.names = FALSE)
```

Final Data Dictionary

variable	class	definition	n_unique	example.value
end_date	Date	last date of Epi week	31	format: 2023-06-03
start_date	Date	first date of Epi week	31	format: 2023-05-28
county	character	county of residence of cases	58	Alameda
health_officer_region	character	California Health Officer Region	6	Bay Area
race_coded	character	race category codes	7	1:07
race_long	character	race category full text	7	Black, Non-Hispanic
race_short	character	race category abbreviated text	7	Black NH
sex	character	sex categorization	2	FEMALE, MALE
age_cat	factor	age category	4	0-17, 18-49, 50-64, 65+
mmwr_week	factor	epi week 40 in 2022 to epi week 23 in 2023	31	22:52
mmwr_year	factor	year	1	2023
cumulative_infected	integer	total number of diagnosed individuals	NA	0 - 137804
cumulative_severe	integer	total number infected requiring hospitalization	NA	0 - 4060
cumulative_unrecovered	integer	total number unrecovered after a week of diagnosis	NA	0 - 16920
new_infections	integer	newly diagnosed individuals	NA	0 - 12110
new_severe	integer	newly identified cases requiring hospitalization	NA	0 - 352
new_unrecovered	integer	newly reported as unrecovered	NA	0 - 1436
pop	integer	estimated population by age group for year 2023	NA	0 - 980387
total_cnty_pop	integer	total population estimate by county	NA	1165 - 9825708
total_race_pop	integer	total county population by race/ethnicity	NA	0 - 4089110
total_sex_pop	integer	total county pouplation by sex	NA	548 - 5049625
total_HOR_pop	integer	total population by Health Officer Region	NA	683715 - 12802429

Descriptive Statistics

Descriptive Statistics: Morbidity Datasets

▼ Code

```
# Bring in final combined database
##-- CA + LA + Population
df <- read.csv(file = here("_data/cleaned_data/combined_df.csv"))

##-- final individual, cleaned databases:
ca_df <- read.csv(file = here("_data/cleaned_data/cleaned_ca_df.csv"))
la_ctny_df <- read.csv(file = here("_data/cleaned_data/cleaned_la_ctny_df.csv"))
pop_df <- read.csv(file = here("_data/cleaned_data/cleaned_pop_df.csv"))
```

to do:

- Overall descriptive stats:
 - infections by age group / county / region / race/ethnicity
 - severe infections by age group / county / region / race/ethnicity
 - unrecovered infections by age group / county / region / race/ethnicity
 - infections/severe infections/unrecovered infections by MMWR week
 - infections/severe infections/unrecovered infections by MMWR week / county or region
- Create a rate metric and summarize to include only one row per strata of interest
 - age-adjusted rates - compare AAR across county / region
 - Incidence rate / densities by age group and month?
 - Incidence rate / densities by race/ethnicity and month?
- Aggregate the data into a new dataframe(s) to include only one row per strata of interest

Calculating Crude Rates

see functions found in: “_functions/calc_total_rates_pop.R”

- add explanation / documentation of functions here and what they do

Infection Rates by Health Officer Region

health_officer_region	total_pop	inf_rate_100k	sev_rate_100k	unrec_rate_100k
Greater Sierra Sacramento	2973210	1184.0	72.3	303.1
Los Angeles	9825708	1402.5	41.3	172.2
Rural North	683715	434.2	31.4	129.9
Central California	4432134	902.8	23.8	100.2
Southern California	12802429	626.2	17.3	73.4
Bay Area	8391874	189.5	11.4	48.2

▼ Code

```
by_HOR_week <-
  counts_week_groups(
    df = df, mmwr_year, mmwr_week,
    health_officer_region
  )

by_HOR_pop <-
  calc_total_rates_pop(
    df = df, pop_col = total_HOR_pop,
    health_officer_region
  )
```

Top 10 Highest Severe Infection Rates by County

county	total_pop	inf_rate_100k	sev_rate_100k	unrec_rate_100k
Inyo	18731	4623.4	427.1	1590.9
Imperial	179639	10494.4	338.5	1364.4
Sierra	3152	2918.8	317.3	1269.0
Alpine	1165	2145.9	257.5	1373.4
Tehama	64677	3110.8	218.0	926.1
Plumas	18955	3001.8	211.0	933.8
Calaveras	44605	2475.1	190.6	784.7
Mariposa	16860	2437.7	189.8	806.6
Modoc	8446	2439.0	189.4	923.5
San Luis Obispo	280251	2158.8	173.4	707.6

▼ Code

```
by_cnty_week <-
  counts_week_groups(
    df = df, mmmwr_year, mmmwr_week,
    county, health_officer_region
  )

by_cnty_pop <-
  calc_total_rates_pop(
    df = df, pop_col = total_cnty_pop,
    health_officer_region, county
  ) %>% arrange(health_officer_region, county)
```

By Sex

- initial note: males have higher infection rate; but females have higher rates of severe infections and unrecovered

sex	total_pop	inf_rate_100k	sev_rate_100k	unrec_rate_100k
FEMALE	849329	15316.8	478.0	1992.2
MALE	806708	17082.3	419.6	1784.0

▼ Code

```
by_sex_cnty <-  
  calc_total_rates_pop(  
    df = df, pop_col = total_sex_pop,  
    health_officer_region, county, sex  
  ) %>% arrange(health_officer_region, county, sex)  
  
by_sex_total <-  
  calc_total_rates_pop(  
    df = df, pop_col = total_sex_pop, sex  
  )
```

By Race and Ethnicity

- initial note: Hispanic population has highest infection rate, severe infection rate, and unRecovered rate – all across the board – this community is being impacted the most.

race_long	total_pop	inf_rate_100k	sev_rate_100k	unrec_rate_100k
Hispanic (any race)	335452	41080.1	1015.3	4295.7
White, Non-Hispanic	485940	11033.5	835.5	3481.9
American Indian or Alaska Native, Non-Hispanic	4569	11775.0	744.1	2626.4
Black, Non-Hispanic	157817	11989.2	621.0	2577.0
Asian, Non-Hispanic	568612	5079.7	345.9	1400.1
Native Hawaiian or Pacific Islander, Non-Hispanic	14710	6233.9	163.2	666.2
Multiracial (two or more of above races), Non-Hispanic	88937	6127.9	133.8	579.1

▼ Code

```
by_race_cnty <-
  calc_total_rates_pop(
    df = df, pop_col = total_race_pop,
    health_officer_region, county, race_long) %>%
  arrange(health_officer_region, county, race_long)

by_race_total <-
  calc_total_rates_pop(df = df, pop_col = total_race_pop, race_long)
```