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).

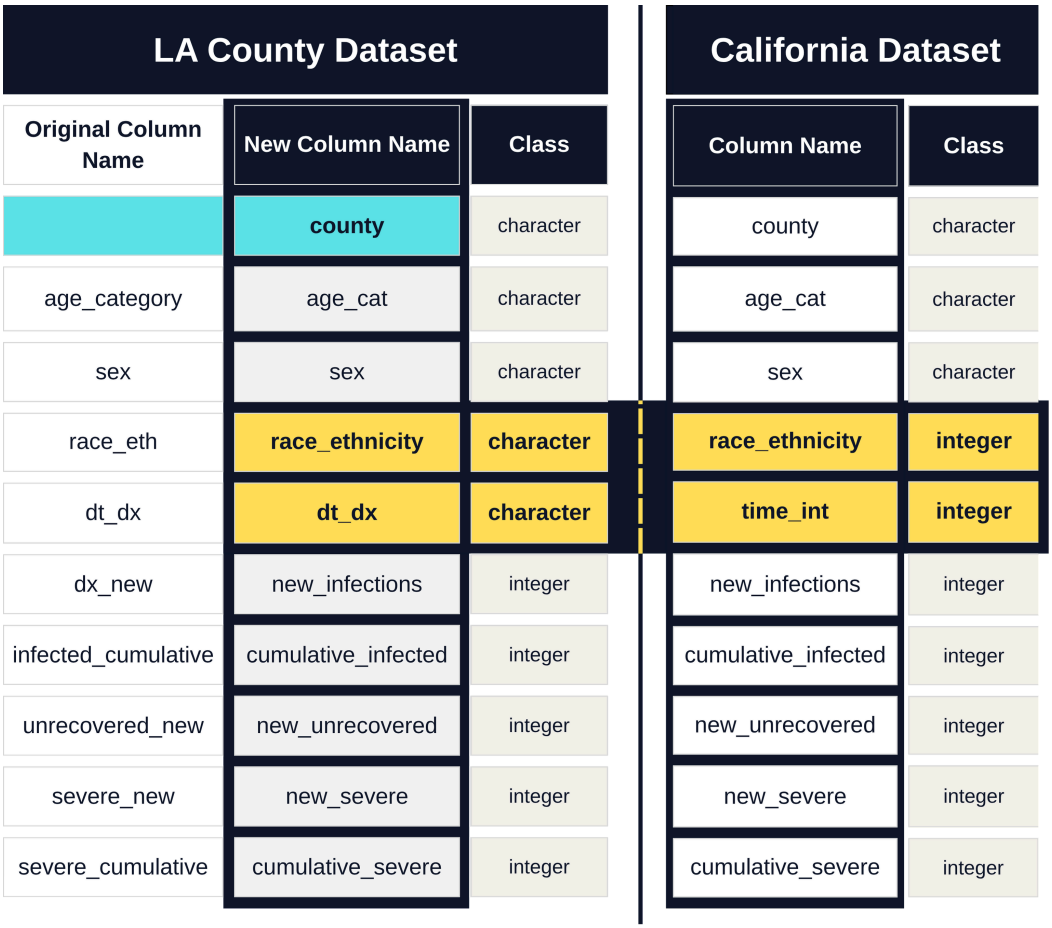
▼ Code

```
##-- 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_cnty_df <- raw_la_cnty_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



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.

▼ Code

```
la_cnty_df <- la_cnty_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())
```

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

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_cnty_df <- la_cnty_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)

pop_df <- pop_df_recat %>%
  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)
```

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

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) %>%
  left_join(pop_df, by = c("county", "sex", "race_coded",
            "race_short", "race_long", "age_cat")) %>%
  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+")))
```

Final Data Dictionary

variable	class	definition	n_unique	examples
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, Alpine, Amador, Butte
health_officer_region	character	California Health Officer Region	6	Bay Area, Greater Sierra Sacramento, Central California, Rural North
race_coded	character	race category codes	7	1, 2, 3, 4, 5, 6, 7
race_long	character	race category full text	7	White, Non-Hispanic, Black, Non-Hispanic, American Indian or Alaska Native
race_short	character	race category abbreviated text	7	WhiteTE NH, Black NH, AIAN NH, Asian 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, 23, 24, 25, 26, 27, 28
mmwr_year	factor	year	1	2023
cumulative_infected	integer	total number of diagnosed individuals	NA	0 - 137804
cumulative_severe	integer	total number of identified individuals having severe disease requiring hospitalization	NA	0 - 4060
cumulative_unrecovered	integer	total number of people reported as unrecovered after a week of being diagnosed	NA	0 - 16920
new_infections	integer	newly diagnosed individuals	NA	0 - 12110
new_severe	integer	newly identified individuals having severe disease requiring hospitalization	NA	0 - 352
new_unrecovered	integer	newly people reported as unrecovered after a week of being diagnosed	NA	0 - 1436
pop	integer	estimated population by age group for year 2023	NA	0 - 980387

Descriptive Statistics