Technical Appendix

All code and data used in the model are publicly available from xxx. This document is intended to complement the codebase by explaining assumptions, discussing modeling decisions, and providing information on data sources.

# Generating microdata

A microdata set from the 2021 ACS was obtained through IPUMS USA with select variables of interest, including county and household identifiers, and sociodemographic variables. Details on how ACS data is collected can be found here. We added a rural/urban flag using county designations defined [here](https://www.counties.org/sites/main/files/file-attachments/2020-june3-countycaucusesinfographic-4-final.pdf). Suburban counties were categorized as rural.

Current asthma prevalence data by age group and county was obtained from the [California Department of Public Health](https://www.cdph.ca.gov/Programs/CCDPHP/DEODC/EHIB/CPE/Pages/CaliforniaBreathingCountyAsthmaProfiles.aspx), sourced from the 2019-2020 California Health Interview Survey. Missing data was imputed using the age-specific California average for age group 0-4 (91% missing), and age-specific population-weighted residual average for age group 5-17 (53% missing). Population weights were derived from the 2010 census, which was the most recent data available.

We recognize that asthma is not evenly distributed within counties. A major factor is proximity to sources of pollution, which are typically in or near lower-income neighborhoods. We therefore used data on poverty (household income as a percentage of the federal poverty line) to adjust the probability of having asthma for each individual in the dataset (Table ##). A binary asthma variable was created using these probabilities.

Table ##. Poverty and asthma prevalence

|  |  |  |
| --- | --- | --- |
| **Household income** | **Proportion of sample** | **Risk ratio for asthma** |
| 200% or more of the poverty line (reference) | 0.75 |  |
| 100% or more but less than 200% of the poverty line | 0.14 | 1.32 |
| Less than 100% of the poverty line | 0.11 | 1.62 |

Lastly, for individuals who were assigned as having asthma, we defined an asthma control variable, by sampling from the vector of asthma control states given the probabilities of having each level of control. These control probabilities were calculated during model calibration and are described elsewhere.

# Data sources

## Census data

The data from which the model population is created was from the 2021 American Community Survey, obtained from [IPUMS USA](https://usa.ipums.org/usa/about.shtml), who provide and maintain a database of public use microdata drawn from numerous federal censuses and national surveys. IPUMS does not identify places with a population fewer than 100,000. As a result, the California data we extracted include individuals from 35 of California’s 58 counties. These 35 counties are representative of 96.4% of the California population. A list of extracted variables is available in the data dictionary available on the project repository.

For urban and rural designation of counties, we used designations provided by the [California State Association of Counties](https://www.counties.org/sites/main/files/file-attachments/2020-june3-countycaucusesinfographic-4-final.pdf). This was operationalized as a binary variable, thus 17 suburban counties were categorized as rural.

## Asthma prevalence

Data on current asthma prevalence was obtained from the [CDPH California Asthma Dashboard](https://www.cdph.ca.gov/Programs/CCDPHP/DEODC/EHIB/CPE/Pages/CaliforniaBreathingCountyAsthmaProfiles.aspx), which reported California Health Interview Survey data, on 6/21/2023. Detailed notes on the data can be found [here](https://www.cdph.ca.gov/Programs/CCDPHP/DEODC/EHIB/CPE/CDPH%20Document%20Library/Notes_About_the_Data_ADA.pdf). The dataset included asthma prevalence by age group and county.

Across all age groups, county-specific prevalence ranged from 0.3% to 33.4%, with an overall California prevalence of 8.7%. Prevalence was mostly higher in suburban and rural counties and was especially high among 5–17-year-olds in Butte County (33.4%), and those over 65 in Solano (23.8%) and Kings (21.4%) counties. (add one of the figures)

As with the census data, prevalence data were suppressed when sample sizes were small in order to protect confidentiality. Younger age groups and smaller counties were therefore missing. The youngest age group of 0-4-year-olds had data for only five counties. We therefore used the California prevalence for that age group (3.3%) when the county-specific prevalence was not available. The next age group of 5-17-year-olds also had 53% missing data. Given greater data availability, we imputed the “residual” average for the missing counties weighted by the population size of 5-17-year-olds in each county. Population size by age were obtained from the 2010 Census, via Social Explorer tables (SE\_T008), to compute county population weights.

## Fire and smoke

We chose the 2018 Camp Fire as given wide availability of evidence for both the smoke pollution it caused as well as health outcomes the event led to. Daily average PM2.5 was selected as the smoke variable. While PM2.5 is among the most dangerous pollutants associated with wildfires, there are limitations to its use as a predictor variable. PM2.5 is not a single substance, rather it refers to a mixture of different chemical compounds. The composition of wildfire PM2.5 greatly differs from ambient pollution and is considered to contain more toxic substances. Measurements of PM2.5 concentration in the air does not provide information about its composition, and therefore does not allow to distinguish differences in potential health impacts. However, the Camp Fire caused substantial increases in PM2.5 concentration in the areas it affected, therefore any PM2.5-related health outcomes over the period of the fire could reasonably be attributed to wildfire smoke rather than ambient pollution.

Daily average PM2.5 measurements were obtained from the [Environmental Protection Agency’s Air Data portal](https://www.epa.gov/outdoor-air-quality-data/download-daily-data). Data was queried daily PM2.5 data for 2018 from all monitors in California. We then subsetted the 15-day period between 8 – 21 November 2018 when the fire was active. The dataset included several zero and negative concentration values, which suggest potential measurement errors. Since so few observations were problematic (0.3%), they were discarded before calculating average daily PM2.5 concentrations for each county. (add barplot)

NOTE::: Do we want to limit to only the counties that were affected from the fire? Could look at the non-fire days average pollution and exclude if no large difference during fire. To ensure we’re capturing health effects of the fire and not ambient pollution – because PM2.5 composition will be different.

**Assumptions:**

Transition probabilities are not time-dependent i.e., they don’t depend on how long people have been in the model and they don’t depend on how long people have been in a given health state.