**TECHNICAL DOCUMENTATION**

* Note regarding data suppression
  + A number of steps have been taken to address data security issue, including 1) aggregation of year into 5-year groups for data displayed at the community and census tract level, 2) showing less granular cause of death data at more granular geographic levels, 3) suppressing all measures for “strata” or “cells” where the corresponding number of deaths is <11, 4) for gender and race/ethnicity stratified data, if one “cell” within a strata is suppressed per #3, *at least one* complementary “cell” is suppressed to avoid arithmetic backcalcuation of the suppressed cell, and 5) excluding disaggregated data by sex for neonatal conditions. These procedures assure compliance with the California Health and Human Services Agency [Data De-Identification Guidelines (DDG)](https://www.dhcs.ca.gov/dataandstats/Documents/DHCS-DDG-V2.0-120116.pdf).
* Note regarding YEAR or Year Group
  + At the County and State levels of geography, YEAR is the individual year of death, with current data from 2001 to 2017. At the Community and Census Tract levels of geography, all data are displayed for the years 2013 to 2017 combined. These years are combined for statistical stability, so that for these more granular levels of geography, the displayed data are still meaningful, and not just the result of random fluctuations.
* Key definitions
  + Communities: Throughout the CCB, communities are defined by Medical Service Study Areas (MSSAs), a unique California geographic designation based on aggregation of census tracts, constructed by the California Office of Statewide Health Planning and Development (OSHPD) with each decennial census [CHHS/OSHPD/MSSA](https://oshpd.maps.arcgis.com/home/item.html?id=a20100c4bf374bd081bb49b82cbaaac3#overview). MSSAs provide the CCB with a good surrogate for “communities” because:
    - 1. there are 542 MSSAs for the 2010 census, providing much more geographic granularity than the 58 California counties and much greater numerical/statistical stability than the 8000+ California 2010 census tracts,
      2. in general, they are aligned with “communities” in the important sense of geographic, cultural, and sociodemographic similarities (although this is generally more true for urban than rural MSSAs, because of the larger size of MSSAs in rural areas),
      3. the names associated with each MSSA have some resonance in many cases with local ideas of “community.”
    - (Although not yet implemented in a fully automated fashion, users can work with the CCB project team to create their own customized communities (based on designated census tracts) for incorporation into the CCB.)
  + Social Determinants of Health: The conditions in which people are born, grow, live, work, and age, including the health system. These circumstances are shaped by the distribution of money, power and resources at global, national and local levels.

**Data and other Key Inputs:**

* Death data
  + Provided by California Department of Health (CDPH), Center for Health Statistics and Informatics (CHSI) [CDPH/CHSI/Death Files](https://www.cdph.ca.gov/Programs/CHSI/Pages/Data-Applications.aspx) (with key information and differences about these files [here)](https://www.cdph.ca.gov/Programs/CHSI/CDPH%20Document%20Library/HIRS-Comparison%20of%20CA%20Death%20Data%20Sources.pdf).
    - Files used: “Death Static Master Files (DSMF)” for 2000 to 2004 and “California Comprehensive Death Files (CCDF)” for 2005 to 2018.
      * Because CCDF Files were used for 2005-2018, deaths of California residents that occurred and were recorded OUTSIDE of California those years have NOT yet been incorporated into any of the CCB working data, visualizations or tables.
    - A death record was considered to be of a California resident based on field “71, Residence State/Province” for the most recent data and on field “46 State of Residence” for 2001-2004 data. A tiny fraction of these records geocoded to locations outside of California, and others had anomalies suggesting the possibility that the residence was not in California. However, the number of such anomalies is relatively minuscule, such that they are extraordinarily unlikely to have any impact on observed patterns and trends.
    - County was based on field “62, Decedent's County of Residence Based on City/State (NCHS Code)” for 2011-2018 data and on field “35, Place of Decedent's Residence” for 2001-2004 data except when modified as noted in “Census Tract Data Issues” below.
    - California death data are geocoded using the CDPH geocoding service, which uses StreetMap Premium for ArcGIS. We have not determined if there is a confidence score or match score below which the census tract for an address is not provided. For 2011-2018, the years where the CCB uses these data to determine census tract and (and therefore communities), a high percentage of records geocoded to a valid census tract (96.4% to 97.2%)?the remaining records contained invalid addresses and/or other anomalies. While this overall rate of geocoding is high, there is substantial variation in the geocoding percent between counties, and some counties, particularly some rural counties have geocoding rates as much as half those noted above
    - Other data coding and cleaning issues:
* Social Determinants of Health (SDOH)
  + The CCB currently contains a small, exploratory set of SDOH variables extracted from the [California Healthy Places Index (HPI)](https://healthyplacesindex.org/) (publicly available files at [HPI](https://healthyplacesindex.org/data-reports/)). The CBD short term road-map includes a plan to extract SDOH data directly from US Census / American Community Survey API (URL) using the [R tidycensus package](https://walkerke.github.io/tidycensus/). Of note, related publicly available data for all census tracts in the United States can be downloaded from the CDC/ASTDR Social Vulnerability Index (SVI) project at [CDC/ASTDR/SVI](https://svi.cdc.gov/data-and-tools-download.html).
* Population data
  + For census tracts (and therefore communities) population denominator data are based on the [American Community Survey](https://www.census.gov/programs-surveys/acs/guidance.html) 5-year extracts (tables B01001\_001E, B01001\_002E, and B01001\_026E) using the most recent 5-year period available corresponding to the 5-year tract/community data being analyzed in the CBD (e.g. 2014-2018 death data uses the 2016 ACS data, which covers 2013-2017). Community population data are generated by aggregating these census data up to the community level.
  + ACS data are extracted directly from the Census/ACS API (Application Program Interface) using the [R tidycensus package](https://walkerke.github.io/tidycensus/).
  + For counties, population denominator data are based on [estimates from the California Department of Finances (DOF)](http://www.dof.ca.gov/Forecasting/Demographics/Estimates/), and are downloaded directly via API from the [State of California Open Data Portal](https://data.ca.gov/dataset/california-population-projection-county-age-gender-and-ethnicity).
* GIS
  + Boundary (or “shape”) files for the CBD were generated using the tracts() function of the [R tigris package](https://github.com/walkerke/tigris), modified to be of smaller file size using the ms\_simplify() function of R rmapshaper package, and with removal of islands off the west coast of some counties using a custom island removal function.
  + Maps do not currently use any explicit projection, but easily could, and probably should, based on user input.
* ICD-10 Mapping
  + In the current version of the CBD project, only the single underlying cause of death ICD-10 code is used. A future release of the CBD may incorporate “multiple cause of death” codes for some conditions.
  + We based the hierarchical list of about 70 disease/injury conditions used in the CBD on a variant of the World Health Organization (WHO) global burden of disease condition list, modified to enhance the usefulness and applicability for U.S. public health priorities and programs. The hierarchy has three levels. The “Top Level” includes “Infectious Diseases”, “Coronary Heart Disease”, “Cancer/Malignant Neoplasms”, “Other Chronic Conditions”, and “Injury” as well as all causes combined. For data displayed at the census tract level, only this level of the hierarchy is included due to sample size and statistical reliability limitations. The next, “Public Health” level, splits each of these top levels into about 50 subcategories, and this is the default level for data/maps displayed at the community level. The final detailed level breaks a few of these Public Health level conditions down further, for the total of about 70 categories. All the levels are shown for data/maps displayed at the county level.
    - County: Top Level, Public Health Level, Detail Level
    - Community: Top Level, Public Health Level
    - Census Tract: Top Level
  + Categorization of deaths was extracted from death certificates based on the International Classification of Diseases version 10 (ICD-10). The primary basis for the ICD10?to-condition mapping was the WHO Annex Table A from “[WHO methods and data sources for global burden of disease estimates 2000-2015, January 2017](http://www.who.int/healthinfo/global_burden_disease/GlobalDALYmethods_2000_2015.pdf)”. We did not use a similar, more recent and more detailed, system developed by the Institute for Health Metrics and Evaluation (IHME) at the University of Washington (recent relevant publications include [The State of US Health, 1990-2016 Burden of Diseases, Injuries, and Risk Factors Among US States, JAMA 2018](https://jamanetwork.com/journals/jama/fullarticle/2678018) and [US County-Level Trends in Mortality Rates for Major Causes of Death, 1980-2014, JAMA 2016](https://jamanetwork.com/journals/jama/fullarticle/2592499)) in this version of the CBD because that system resulted in 721,783 (19.2%) of California deaths from 2000 to 2015 being mapped to “garbage codes”, for which more sophisticated methods would need to be employed. The possibility of redistributing these “garbage codes” to valid categories at the census tract level and otherwise using the IHME system is being explored and may be implemented in future versions of the CBD. However, to enhance our use of the WHO system we compared the mapping of 3,758,856 deaths based on the WHO and IHME systems and changed the WHO mapping of ICD codes for several categories wherein the IHME classification was considered more appropriate (e.g., specific cancer sites rather than “other malignant neoplasms.”). All of these modifications are carefully described in a key resources tool for the CBD, available [here](https://github.com/mcSamuelDataSci/CACommunityBurden/blob/master/myCBD/myInfo/gbd.ICD.Map.xlsx) on our GitHub site. In addition, because our focus was on the “Public health” list of conditions, we remapped a number of ICD-10 codes from the WHO mapping to our own CBD system. All of these modifications are documented in a “key resources” tab for the CBD available noted above.
  + The latest IHME/GBD results and methods can be found [here](https://www.thelancet.com/journals/lancet/issue/vol392no10159/PIIS0140-6736(18)X0048-8)
* Census Tract Data Issues
  + Of the 8041 California census tracts in the 2010 Census “Tiger” files, five have zero land area (only water), and (not surprisingly…) have zero population and zero deaths. These tracts are excluded from CCB data processing and display.
  + Of the remaining 8036 tracts, 12 have zero population and zero deaths. By definition, these tracts are excluded from all analyses, and will show values of “0” or missing for all measures in all maps. These tracts are wholly comprised of industrial facilities, airports, or parks.
  + Census tract issues to be addressed:
    - Census tracts (and communities) where greater than 50 percent of the population live in congregant living quarters will noted in the future with an “\*” on relevant maps and charts in an upcoming CCB release. For some comparisons (e.g. of rates) these tracts could be removed from the larger geographies in which they are contained.
    - During detailed review of multiple data sources, we observed a number of instances where stated county of residence was not consistent with the census tract to which that death was geocoded. In these instances we will consider recoded the county based on the address and subsequent geocode.

**Formulas and Measures**

* Years of Life Lost (YLL)
  + Two methods are available for calculating YLL.
  + In the current implementation of the CCB, we use the first method below. For local health departments, or others using their own implementation of the CBD, both method are coded, and either one can be selected (in the “E1.make\_death\_datasets\_for\_app.R” file)
    - The first method, is simpler, and is based on summing for *all* deaths, the number of years prior to age 75 that *each* death occurs, with 0 YLL used for deaths occurring at ages >= 75. This method has the advantage of (1) emphasizing more strongly deaths that occur at younger ages and (2) being simpler to explain and understand. It has the disadvantage of not being consistent with the methods of the Institute for Health Metrics and Evaluation, prohibiting direct comparisons with their results.
    - The second method uses the current approach of the Global Burden of Disease Study. With this method the YLL for each death is based on the age at death, and the additional number of years a person living in an optimal setting could be expected to live (page 30, [here](http://www.who.int/healthinfo/global_burden_disease/GlobalDALYmethods_2000_2015.pdf)). For example, someone dying at birth would be associated with 91.94 YLL, someone dying at 25 associated with 67.08 years, and someone dying at 98 with 3.70 years. Beyond the published data, when we use this method, we associate 1.0 YLL for anyone dying above age 105.
      * Our mapping of age at death to YLL for this method can be found on our GitHub site [here](https://github.com/mcSamuelDataSci/CACommunityBurden/blob/master/myCBD/myInfo/le.Map.xlsx).
* Crude rates
  + All rates are expressed per 100,000 people based on the following calculations:
    - 100,000\*(number (e.g. deaths, potential years of life lost) / midyear population)
  + Confidence intervals for crude rates are based on the pois.approx() function of the [R epitools package](https://github.com/cran/epitools).
* Age adjusted rates
  + Age-adjusted rates are based on the “direct” method, using standard definitions and procedures. Great descriptions and the motivations for these methods can be found [here](https://www26.state.nj.us/doh-shad/sharedstatic/AgeAdjustedDeathRate.pdf), from the New Jersey Department of Health, and [here](https://www.cdc.gov/nchs/data/statnt/statnt06rv.pdf), from CDC.
  + The US 2000 Standard Population from [NCI](https://seer.cancer.gov/stdpopulations/) and [CDC/NCHS](https://www.cdc.gov/nchs/data/statnt/statnt20.pdf) was used; details of the methods and implications of using the 2000 standard population are described by NCHS and can be found [here](https://www.cdc.gov/nchs/data/nvsr/nvsr47/nvs47_03.pdf)
  + Ten age-groupings were used for these calculations.These groups and the corresponding standard population data can be found [here](https://github.com/mcSamuelDataSci/CACommunityBurden/blob/master/myCBD/myInfo/Age%20Groups%20and%20Standard%20US%202000%20pop.xlsx).
  + The age-adjusted calculation, and generation of confidence intervals was conducted using the “ageAdjust.Direct()” function of the [R epitools package](https://github.com/cran/epitools).
  + Because a very small number of census tracts with otherwise useful data had zero population in one or more age strata (often the youngest or oldest strata, for just one sex), the above-mentioned function was modified such that rates in such strata were assigned to (reasonably enough) be 0 (rather than undefined/infinity), allowing an adjusted rate to be calculated.
* Life Expectancy
  + Life tables for tracts, communities, counties and states are generated from age specific mortality rates, which are the quotient of deaths during a calendar year to the and exposure , approximated by the population of the same age at the midpoint of the year (July 1). Age structured population data for tracts and communities are estimated using data from the American Community Survey, 5-year sample (table B01001; multiple years). County and state age population by age are estimated by the Demographic Research Unit, CA Department of Finance. Deaths data are based on 100% extracts from the vital statistics reporting system, CA Department of Public Health.
  + Life tables are estimated for tracts in counties where at least 95% of deaths could be accurately geocoded to the tract level. Mortality and exposure data were combined for small geographies: 4 years? historical data were added to the population and mortality calculations for each annual tract and community life table, and 2 years? historical data were added to each county life table. Then, life tables were calculated for geographies meeting a minimum of at least 10,000 person-years of exposure. Intra-age mortality (nax) was calculated for ages below 5 using factors provided by Preston et al. (2001) and by the midpoint of the age interval for other age groups. Standard errors were calculated for age specific probabilities of death (Chiang 1984) and simulated life tables were generated by bootstrapping life table deaths to produce confidence intervals for life expectancy (Andreev & Shkolnikov 2010).
    - Preston, Samuel H. and P. Heuveline and M. Guillot. 2001. Demography. Blackwell, pp. 47-48.
    - Chiang, C.L. 1984. The Life Table and its Applications. Robert E Krieger Publ Co., pp. 153-168.
    - Andreev, E.M. and V. M. Shkolnikov. 2010. ?Spreadsheet for calculation of confidence limits for any life table or healthy-life table quantity.? Max Planck Institute for Demographic Research: MPIDR Technical Report 2010-005; June 2010.

**Hospitalization Information**

* Data source
  + Hospitalization data are based on 2016 nonpublic Patient Discharge Data received from the [California Office of Statewide Health Planning and Development (OSHPD)](https://oshpd.ca.gov). OSHPD provides such files from inpatient data they collect from California-licensed hospitals. The data set consists of a record for each inpatient discharged from a California-licensed hospital. Licensed hospitals include general acute care, acute psychiatric, chemical dependency recovery, and psychiatric health facilities. Data are not collected from Veteran’s Administration, Military or other Federal Hospitals or from Tribal Hospitals.
  + Detailed information for the current OSHPD Patient Discharge Data and data system can be found [here](https://oshpd.ca.gov/ml/v1/resources/document?rs:path=/Data-And-Reports/Documents/Request/Data-Documentation/DataDictionary_PDD_2018_Nonpublic.pdf) and a link to the 2016-specific data can be found [here](https://oshpd.ca.gov/data-and-reports/request-data/tools-resources/data-documentation/).
* Charges
  + For each hospitalization one total charge is listed, reflecting the charges associated with the primary condition as well as any other charge associated with the hospitalization.
    - Total charge is all charges for services rendered during the length of stay for patient care at the facility, based on the hospital’s full established rates. Charges include, but are not limited to daily hospital services, ancillary services, and patient care services.
    - Hospital-based physician fees are excluded, as are items like take-home drugs, television, follow-up home health visits, ambulance services, etc.
    - Hospitals report ‘Total Charges’ to OSHPD for the last 365 days of stay. However, starting in 2015, in the files released by OSHPD, ‘Total Charges’ are adjusted to reflect the entire length of stay.
    - A hospitalization can have multiple discharges if the patient moves between type of care (e.g. psychiatric to acute would be two discharges during the same overall hospital stay) and each discharge has a related total charge.
    - Charges of $1 specify ‘no charge’ or charity care.
  + The noted charges are based on hospitals’ administrative systems and do not indicate actual costs/payments for those charges.
  + Nevertheless, because the CCB describes summary data, the charts and tables shown provide valuable information regarding the patterns of the monetary burden of disease/conditions in California from the hospitalization perspective.
  + For some hospitalizations, no charges are reported to OSHPD, and for some hospitalizations implausibly high charges (likely errors) have been excluded, so total charges may be slight underestimates from this perspective. ‘Average’ charges in these charts are based on the median rather than the mean, so are largely not impacted by these issues.
* ICD-10-CM Codes
  + For each hospitalization, one condition is established and coded as the chief cause of the admission, and is noted as the Principal or Primary diagnosis. Up to 24 other conditions that coexist at the time of admission, that develop subsequently during the hospital stay, or that affect the treatment received are also included and noted as Other or Secondary diagnoses.
  + Coding for these Principal and Other diagnoses are based on the ICD10-CM system (from October 1, 2015 forward; prior to this time ICD9-CM was used), along with standardized guidance.
  + The codes entered by the hospitals are subject to multiple sources of variation or potential error (e.g. selective use of codes for billing purposes). Nevertheless, since the data are used in the CCB in summary form, the overall patterns displayed are likely to be meaningful and informative.
* Grouping of ICD-10-CM Codes
  + ICD10-CM codes are highly detailed and specific, with more than 93,000 codes. There are many ways these codes can be grouped or summarized into meaningful categories. We continue to explore which grouping system(s) are most useful for purposes of the CCB, and would welcome your input. At this time, settings are available to toggle between these options. Four possible systems include:
    - The Global Burden of Disease system (GBD) system (coded by the CCB team), groups the hospitalization ICD-10-CM codes into conditions based on, generally, the Global Burden of Disease system, as described above for death data, and includes ‘high volume’ conditions and some other conditions of clear programmatic public health interest in California. At the moment, this system is not exhaustive, and includes only a sample of conditions of interest.
    - The Major Diagnostic Categories (MDCs) system (included in the OSHPD data set), groups principal diagnoses into 25 mutually exclusive diagnosis groupings. The diagnoses in each MDC correspond to a single organ system or etiology and, in general, are associated with a particular medical specialty. The system was established and is maintained by the U.S. Department of Health and Human Services (DHHS) Centers for Medicare and Medicaid Services (CMS).
    - The Medicare Severity Diagnosis Related Group (MS-DRG) system (included in the OSHPD data set), categorizes patients based on clinical coherence and expected resource intensity, with respect to diagnoses, treatment and length of hospital stay. The assignment of an MS-DRG is based on: the principal diagnosis and any secondary diagnoses, surgical procedures performed, comorbidities and complications, patient’s age and sex, and discharge status. The system was established and is revised annually by CMS. See CCR Section 97212.
    - The [Clinical Classifications Software (CCS)](https://www.hcup-us.ahrq.gov/tools_software.jsp) system (included in the OSHPD data set) is a tool provided by the Agency for Research and Quality Healthcare Cost and Utilization Project. CCS aggregates the ICD codes into a manageable number (285) of clinically meaningful categories to make it easier to quickly understand diagnosis patterns. The system is evolving, with the current system organized across 21 body systems, which generally follow the structure of the ICD-10-CM diagnosis chapters.