* Key definitions
  + “Communities”—throughout the CCB, communities are defined by “Medical Service Study Areas”, a unique California geographic designation, based on aggregation of census tracts, constructed by the California Office of Statewide 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, and (2), as much as possible, they are aligned with “communities” in the important sense of geographic, cultural, and sociodemographic similarities—this is generally more true for urban then rural MSSAs, because of the larger size of MSSAs in rural areas, and (3) the names associated with each MSSA has some resonance in many cases with “community” as noted above. Although not yet implement in the CBD in a fully automated fashion, users can work with the CCB project team to provider their own list of communities—based on designated census tracts---and these can easily be brought into a local, or even statewide implementation of the CCB.
  + Social Determinants of Health: [need short prose here]
* 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).
      * “DSMF” files for 2000 to 2004 and “CCDF” files for 2005-2015. Information on files [here](https://www.cdph.ca.gov/Programs/CHSI/CDPH%20Document%20Library/HIRS-Comparison%20of%20CA%20Death%20Data%20Sources.pdf).
      * A death record was considered to be of a California resident based on field “71, RESIDENCE STATE/PROVINCE” for 2005-2015 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 California. Efforts are ongoing to address these issues. However, the number of such anomalies is so relatively minuscule, that they are extraordinarily unlikely to have any impact on observed patterns and trends.
      * County was based on field “62, Final Decedent's County of Residence Based on City/State (NCHS Code)” for 2011-2015 data and on field “35, Place of Decedent's Residence” for 2001-2004 data; except as where modified as noted in “Census Tract Data Issues” below.
      * Prior to CCB project receipt of these death data, CHSI staff geocoding the data using ….. For the years where the CCB uses these data for census tract and (and therefore community) geographies, 2011-2015, a high percent of records geocoded to a valid census tract (9X.X% to 9X.X%)—the remaining records contained invalid addresses and/or other anomalies.
      * Other data coding and cleaning issues:
  + Social Determinants of Health (SDOH)
    - The CCB currently contains a small, exploratory set of SDOH variables extracted from [California Healthy Places Index (HPI)](https://healthyplacesindex.org/)  publically 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/SVIDataToolsDownload.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 5-year period corresponding to the 5-year tract/community data being analyzed in the CBD (i.e. currently 2011-2015 death data therefore using the 2015 ACS data, which covers 2011-2015). 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 [CA.gov Open Data Portal](https://data.ca.gov/) here [CHHS/DOF/Open Data](https://data.ca.gov/dataset/california-population-projection-county-age-gender-and-ethnicity).
* GIS
  + Boundary (or “shape”) files for the CBD were generated (as “simple features” objects, using the R sf package) using the tracts() function of the [R tigris package](https://github.com/walkerke/tigris), modified to be smaller using the ms\_simplify() function of R rmapshaper package, and with removal of physical islands of the west coast of some counties using a custom xxx function.
  + Maps are projected based on the xxx projection (although this can be easily changed as needed).
* ICD-10 Mapping
  + In the current version of the CBD project, only the single underlying cause of death ICD-10 code is use to classify the cause of death. Efforts are underway to incorporate ‘multiple cause of death” codes for some conditions in a future release of the CBD, and based on user interest/support.
  + We used a variate of the WHO/GBD condition list hierarchical outline to create our own hierarchical outline with three levels. The first 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 hierarch is included. The next level, our “Public Health Level” breaks each of each of the top level down into a total grouping of 51conditions; the final detailed level breaks a few of these PH level conditions down further, for a total of 66 conditions, which is displayed at the county level.
  + The primary basis for the ICD-10 to condition mapping is the WHO/GBD 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)”. For the primary mapping, we were not able to use the IHME/GBD mapping system as described in the Supplemental Appendices to their recent publications (e.g. [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)) because that system results in approximately XX% of (California) death being mapped to “garbage codes”, for which more sophisticated, and possibly not appropriate for our system, methods would need to be employed. However, to enhance or use of the WHO/GBD system we did compared the mapping of XX,XXX,XXX California deaths based the two systems. As a result of these comparisons we did find a number of instances where the two systems deviated meaningfully in their mappings, and for which we believed we could improve the WHO/GBD “mapping” by using the IHME/GBD mapping. All of these modifications are carefully described in a key resources tool for the CBD, available HEREXXX on our GitHub site.
  + In addition, because of our focus on the “Public health” list of conditions, we remapped a number of ICD-10 codes from the GBD mapping to our own CBD system. These include alcohol… All of these modifications are also carefully described at the site above.
* Census Tract Data Issues
  + xxx census tracts included in some designations of California 2010 census tracts were excluded for all project calculations and maps for communities or census tracts because they are listed as containing 0 population.
  + Another xxx tracts with population less than xxx were also excluded because of the limited utility of any data from these tracts.
  + Census tracts (and communities) where greater than X percent of the population live in congregant living quarters ARE/WILL BE noted with an “\*” on relevant maps and charts. For some comparisons (e.g. of rates) these tracts could be removed from the larger geographies in which they are contained, based on user request.
  + Based on 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 geocode—in these instances we “recoded” the county based on the addressed and subsequent geocode.
* Formulas and measures
  + Years of Life Lost (YLL)
    - Following the methods of the Global Burden of Disease Study, 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, we associated 1.0 YLL for anyone dying above age 105.
    - Our mapping of age at death to YLL 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 as per 100,000 population based on 100,000\*(number of events (e.g. deaths) / 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 the standards definitions and procedures. Great descriptions and the motivations for these methods can found [here](https://www26.state.nj.us/doh-shad/sharedstatic/AgeAdjustedDeathRate.pdf), from the New Jersey Department of Health.
    - Ten age-grouping were used for these calculations
    - 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.
    - The age group mapping at this standard population data can be found here URL.
    - 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 0 population in one or more age strata (often the youngest or oldest strata, often 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
* Other R tools and packages