Institute for Health Metrics and Evaluation

Methods overview for the Get With The Guidelines stroke data linkage project

Prepared by the *Institute for Health Metrics and Evaluation*
for the *American Heart Association*

December 2022

Megan Lindstrom, PhD

Feras Wahab, MS

Kate LeGrand, MPH

Greg Roth, MD

Executive summary

The purpose of this project is to enrich a deidentified extract from the AHA GWTG-Stroke Registry with IHME's county-level health and health-related data. The addition of county-level data to this registry will support the AHA's goal of increasing the availability of population-level variables for analyses performed with this registry data. The overall goal is to make a merged data table available on the AHA's Precision Medicine Platform (PMP) for use by research scientists in a data challenge hosted by AHA.

The AHA GWTG-Stroke Registry is a national registry of hospitalized stroke cases. This registry is used to create a deidentified, person-level dataset for research purposes.

IHME county-level data is the result of work by IHME's US Health Disparities team and includes US county level life expectancy, all-cause mortality, and for selected causes the cause-specific mortality and cause-specific years of life lost, as well as sociodemographic data such as education and income levels. For selected variables, this county-level data is available by 5-year age strata, sex, and year for years 2010 through 2019.

IHME will create a merged data table that combines the AHA GWTG-Stroke Registry with key population characteristics and burden of disease variables for the US available from IHME's Global Burden of Disease Study.

This documentation package introduces the input data and analytic methods that will be used to produce the final merged data table. We present here descriptions of how IHME generates county level health metrics, the geographical transformations of data to include county and zip code processing, and how we will merge IHME data with the GWTG-Stroke registry extract. The material in this package may be used to guide the development of the research questions for the AHA data challenge. Accompanying the final data table will be a comprehensive documentation package, including data dictionaries and codebooks, for participating researchers.

Health metrics

IHME will produce county-level estimates for life expectancy, all-cause mortality, total stroke mortality, and mortality due to the following stroke subtypes: ischemic stroke, intracerebral hemorrhage, and subarachnoid hemorrhage for the years 2010-2019. These results will be made available by total population, age, sex, and race/ethnicity.

These mortality estimates will be accompanied by county-level sociodemographic covariates related to education and income across the same time series. IHME has produced the following covariates, many of which are also available by race/ethnicity:

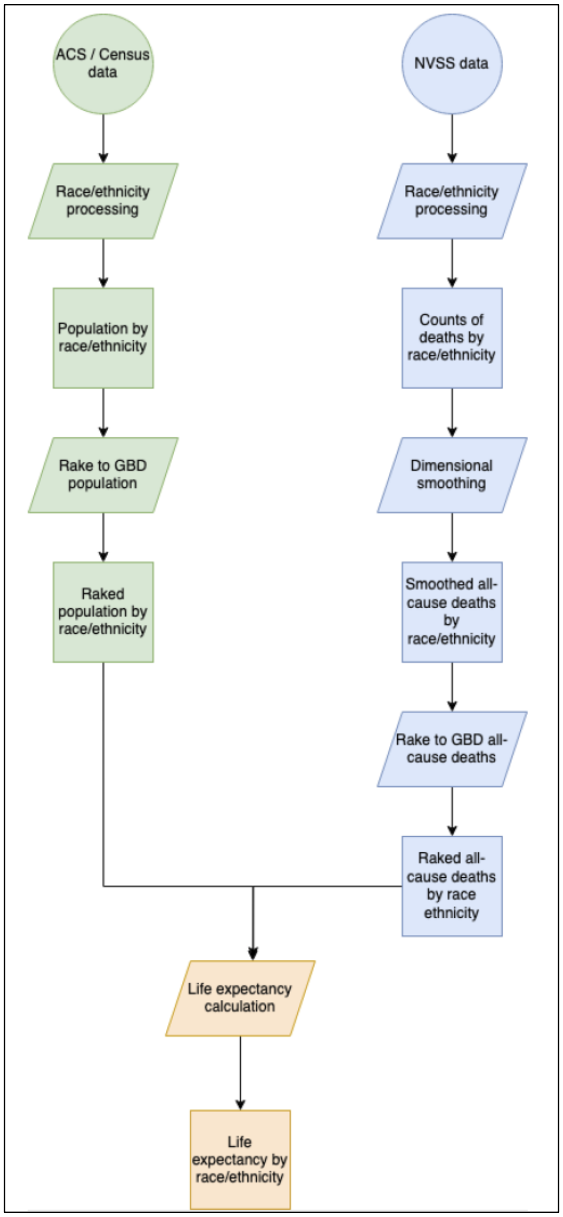
- Proportion of population graduated from college
- Proportion of population graduated from high school
- Per capita income
- Poverty
- Unemployment
- Median household income
- Foreign-born status
- Home ownership

Life expectancy estimation methods

Death certificate and census data were combined to create smoothed time series for each race/ethnicity group in each state (Johnson et al., 2022). We used bridged race/ethnicity data from the U.S. Census to construct a time series by age, sex, and state–race/ethnicity group from 1990 to 2019 (Centers for Disease Control and Prevention). To generate estimates of all-cause mortality, we used death certificate data from the National Vital Statistics System. To generate a complete time series from 1990 to 2019, we modeled counts of deaths using a structured linear regression with a multidimensional Gaussian smoother to borrow strength across multiple dimensions (age and time), informed by observed residuals and their uncertainty. To carry the uncertainty from the models through the rest of the process, we took 1000 draws from the posterior distribution of the model.

To generate estimates of LE at birth, we constructed bridged period life tables, using the draws of all-cause mortality rates described in the previous paragraph as input data. This approach generated 1000 draws of LE for each age, sex, year, and state–race/ethnicity group. For each group, point estimates were generated by taking the mean of the draws, and the bounds of the 95% uncertainty intervals (UIs) were determined by taking the 2.5th and 97.5th percentiles of the draws. Supported changes in LE were interpreted as a 95% UI for percentage change that did not cross 0. To capture differences between leading and lagging locations, absolute disparities in LE between state–race/ethnicity groups were calculated as the difference between LE point estimates. Some state–race/ethnicity groups in some years had very few deaths, most notably for the oldest age groups among smaller populations. We excluded those state–race/ethnicity–location groups when the average number of deaths for either sex in any decade was fewer than 10 for the terminal age group of 85 years or older.

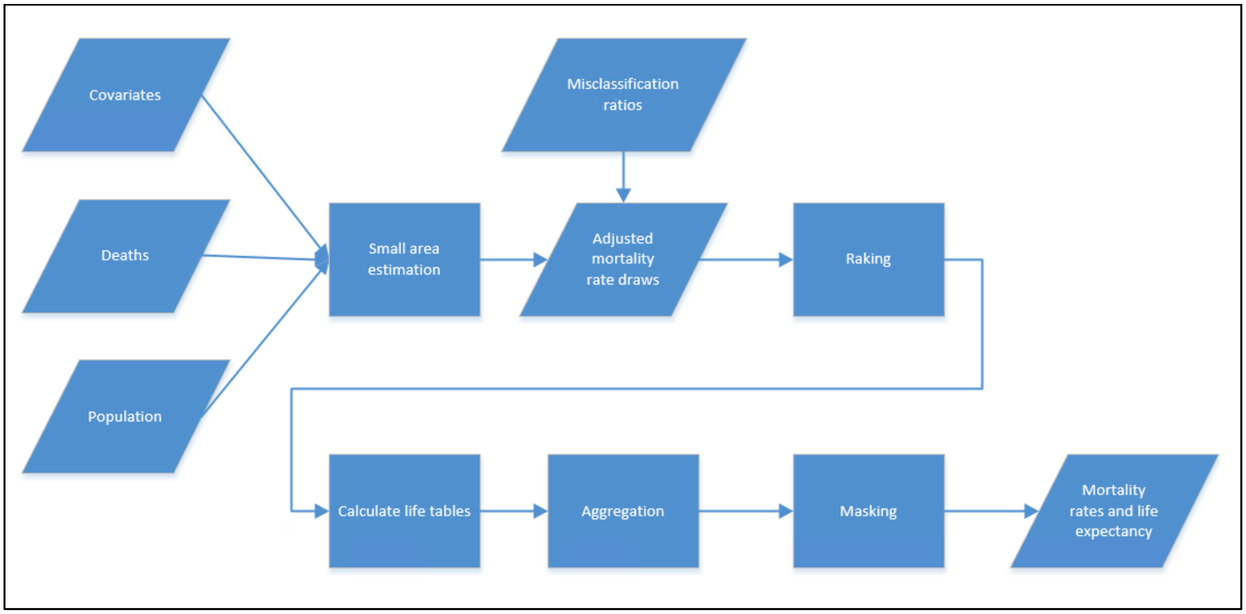
The flowchart below depicts the life expectancy estimation process:


ACS / Census data
NVSS data
Race/ethnicity processing
Race/ethnicity processing
Population by race/ethnicity
Counts of deaths by race/ethnicity
Rake to GBD population
Dimensional smoothing
Raked population by race/ethnicity
Smoothed all-cause deaths by race/ethnicity
Rake to GBD all-cause deaths
Raked all-cause deaths by race ethnicity
Life expectancy calculation
Life expectancy by race/ethnicity

Fatal estimation methods

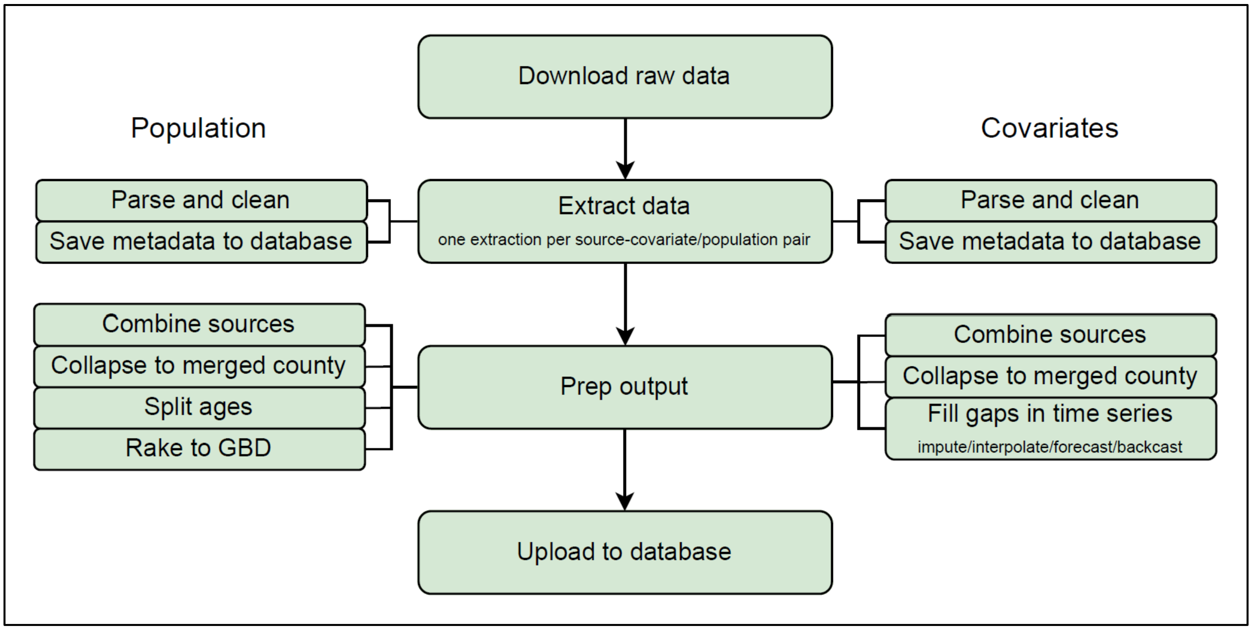
IHME produced the mortality estimates used in this dataset. The estimates are stratified by age, sex, year, geography, and either race/ethnicity. De-identified death records from the National Vital Statistics System and population data from the National Center for Health Statistics were used to build the county-level mortality estimates.

The first step to create the estimates was small area estimation models using the Template Model Builder (TMB) package in R, which estimate cause-specific mortality rates by cause, county, racial/ethnic group, sex, age, and year. TMB utilizes an empirical Bayes approach, leveraging Laplace approximations to estimate the posterior distribution. The second step was to use race/ethnicity misclassification ratios to adjust the results from the small area models. Misclassification ratios are the ratio of deaths among individuals among a particular racial/ethnic group from self-reported deaths to deaths among individuals of a particular racial/ethnic group from death certificates. This was done to adjust for known biases that exist in racial/ethnic reporting on death certificates. Next, a post-hoc calibration on each of the adjusted results was completed to guarantee internal consistency between all causes, geographic levels, and strata. This was completed with a raking algorithm, which proportionally scales the rows to add up to their marginal totals, then scaling the columns the same way until the entries in the table stabilize. The final steps include creating total estimates by combining the estimates for sex at aggregated geographic levels (state and national) through population weighted age-specific mortality rates and masking mortality rates in areas that had a mean annual population of less than 1,000.



Covariate estimation methods

IHME produced a series of sociodemographic covariates for this dataset. The underlying raw data is primarily from the American Community Survey (ACS) and the decennial population census. For instances where the stratified raw data contained missing or unstable estimates due to small populations, the data needed to be completed to provide covariate estimates for each year, location, and strata. Small area imputation models in the form of Bayesian binomial likelihood imputation models were used in the R-INLA package to create smooth covariate estimates in all years, locations and racial/ethnic groups. Data that did not exhibit instability was left in an unimputed format. Both were then uploaded to the dataset.


Download raw data
Population
Covariates
Parse and clean
Save metadata to database
Extract data
one extraction per source-covariate/population pair
Parse and clean
Save metadata to database
Combine sources
Collapse to merged county
Split ages
Rake to GBD
Prep output
Combine sources
Collapse to merged county
Fill gaps in time series
impute/interpolate/forecast/backcast
Upload to database

Merging with geographic data and GWTG-Stroke registry

Linking zip codes to counties

Zip code data are not defined geographic regions, instead they are a collection of routes defined by the United State Postal Service (USPS) which can cross county lines. Due to this, additional steps (beyond simply assigning one zip code to one county) must be completed to move between the two geographies. The first step is to retrieve data which assigns every zip code to every county they fall within. This was retrieved from the United States Housing and Urban Development (HUD) (Office of Policy Development and Research). HUD provides a yearly file for each quarter from 2010-2021, which shows the counties that each zip code falls within. This is updated quarterly due to the frequent route changes instituted by the USPS. We chose to use the fourth quarter file for each year as it would capture all changes that occurred within the calendar year. The file also contains the ratio of residential, business, or other addresses that fall within each county per zip code, which we used to remove duplicate zip code to county records. We chose to keep the zip code to county record that had the highest residential to business or other address type ratio as it the county the patient would most likely live within. The final file merged each of the yearly files, resulting in ten years of data (2010-2019) that links each zip code within the United States to the County they fall within.

Limitations

The first limitation is inherent in patient supplied address data. Some patients may use an address that is temporary (parents addresses, shelters, etc.), which could potentially lead to a misassignment of their home county. Further, the patient home zip code data was incomplete in a significant percentage of records which could create further bias.

The second limitation is the assignment of zip codes to the county that has the most residential addresses compared to other types of addresses. It is possible that the patient could have been located within the other county (or counties) that the zip code fell within, however by using the ratio we provide the most likely county of residence.

Merging with GWTG-Stroke registry

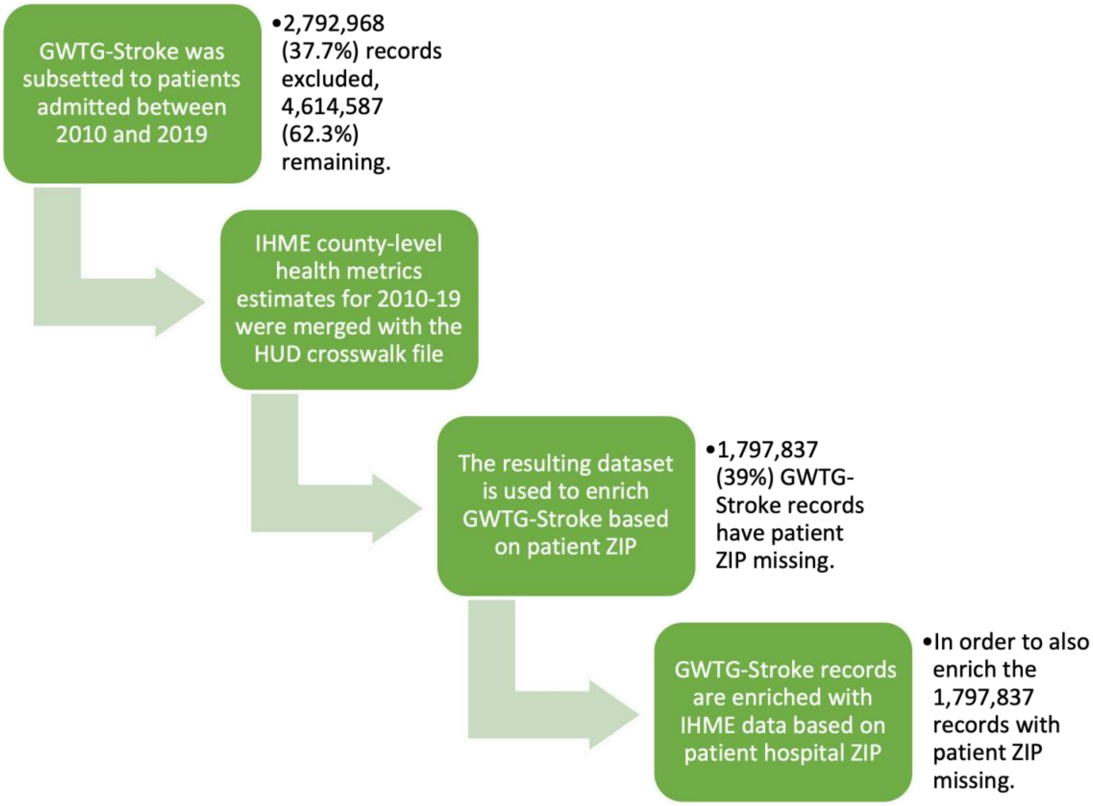
The AHA GWTG-Stroke Registry is a national registry of hospitalized stroke cases. This registry is used to create a deidentified, person-level dataset for research purposes.

The following steps will be taken to merge IHME estimates with the GWTG-Stroke registry:

1) IHME county-level health metrics estimates for the years 2010-2019 were merged with the HUD crosswalk file (HUD2010_2021.csv), in order for each county to be merged with its corresponding ZIP code.
2) GWTG-Stroke data was limited to cases with an admission date of 2010-2019, as indicated by the AdmitYr variable.
3) The resulting GWTG-Stroke dataset was enriched with IHME data based on patient zip code (GS_ZIP1 in GWTG-Stroke and ZIP in HUD2010_2021.csv) and admit year (AdmitYr in GWTG-Stroke and year in HUD2010_2021.csv).
4) As GS_ZIP1 in GWTG-Stroke had a high rate of missingness (39%), patient records were also enriched with IHME health metrics data based on patient hospital ZIP codes (SITE_POSTAL_CODE in GWTG-Stroke and ZIP in HUD2010_2021.csv) and admit year.

Merging GWTG-Stroke with HUD Data

The schematic below describes the process by which GWTG-Stroke records will be enriched with IHME's county-level health metrics data, using HUD data as a crosswalk between ZIP codes and county estimates:



GWTG-Stroke was subsetted to patients admitted between 2010 and 2019

- 2,792,968 (37.7%) records excluded, 4,614,587 (62.3%) remaining.

IHME county-level health metrics estimates for 2010-19 were merged with the HUD crosswalk file

The resulting dataset is used to enrich GWTG-Stroke based on patient ZIP

- 1,797,837 (39%) GWTG-Stroke records have patient ZIP missing.

GWTG-Stroke records are enriched with IHME data based on patient hospital ZIP

- In order to also enrich the 1,797,837 records with patient ZIP missing.

References

Johnson, C. O., Boon-Dooley, A. S., DeCleene, N. K., Henny, K. F., Blacker, B. F., Anderson, J. A., Afshin, A., Aravkin, A., Cunningham, M. W., Dieleman, J. L., Feldman, R. G., Gakidou, E., Mokdad, A. H., Naghavi, M., Spencer, C. N., Whisnant, J. L., York, H. W., Zende, R. R., Zheng, P., … Roth, G. A. (2022). Life expectancy for white, black, and Hispanic race/ethnicity in U.S. states: Trends and disparities, 1990 to 2019. *Annals of Internal Medicine*, *175*(8), 1057–1064. https://doi.org/10.7326/m21-3956

Centers for Disease Control and Prevention. (2022, October 28). *U.S. Census populations with bridged race categories*. Centers for Disease Control and Prevention. Retrieved December 14, 2022, from http://www.cdc.gov/nchs/nvss/bridged_race.htm

Office of Policy Development and Research (PD&R). (n.d.). *HUD USPS ZIP Code Crosswalk files*. HUD USPS ZIP Code Crosswalk Files. Retrieved December 14, 2022, from https://www.huduser.gov/portal/datasets/usps_crosswalk.html