Gather Appendix

5. Provide information on all included data sources and their main characteristics. For each data source used, report reference information or contact name/institution, population represented, data collection method, year(s) of data collection, sex and age range, diagnostic criteria or measurement method, and sample size, as relevant.

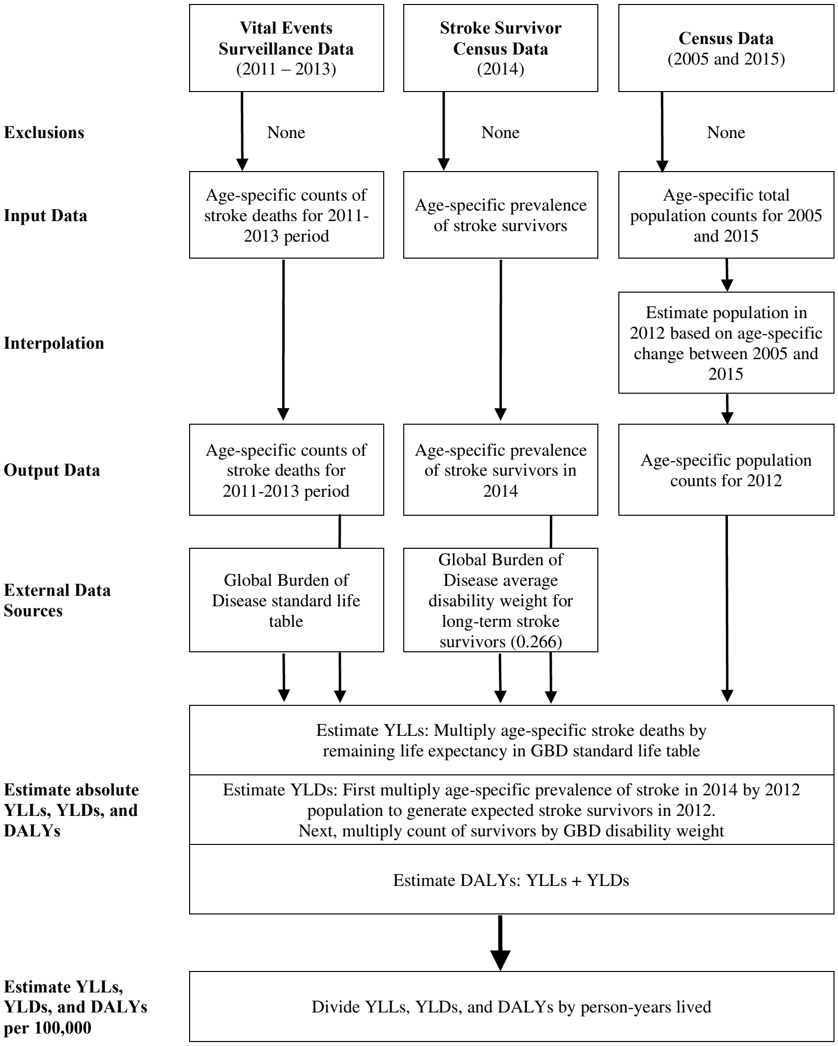
Data Source	**Collecting Institution**	**Population Represented**	**Data collection method**	**Years of data collection**	**Age and sex representation**
Gadchiroli surveillance site population census	SEARCH	All individuals living in the 86 villages that form the Gadchiroli surveillance site	House to house enumeration in 2005 and 2015.	2005 and 2015	Both sexes and all ages
Gadchiroli vital event registry	SEARCH	All individuals living in the 86 villages that form the Gadchiroli surveillance site	All births and deaths are reported to the research team by the community health workers of SEARCH who are residents of the village. Annual house to house cross-survey is conducted to find out any missing births and deaths.	Continuously between 2011 and 2013	Both sexes and all ages
Gadchiroli stroke prevalence data	SEARCH	All individuals living in 39 of the 86 villages that form the Gadchiroli surveillance site	We conducted a three-stage screening. First, a house to house survey of all individuals living in 39 villages was done by trained surveyors using a validated screening questionnaire. Those who	2014	Both sexes and all ages

			screened positive were evaluated by a trained physician in the second phase. Doubtful cases were evaluated by the study neurologist in the third phase.		

8. Provide all data inputs in a file format from which data can be efficiently extracted, including all relevant meta-data listed in item 5. For any data inputs that cannot be shared because of ethical or legal reasons, such as third-party ownership, provide a contact name or the name of the institution that retains the right to the data.

Input data could not be shared due to ethical reasons as the consent of the participant at the time of data collection did not include a clause to share deidentified data with third parties.

9. Provide a conceptual overview of the data analysis method. A diagram may be helpful.


Vital Events Surveillance Data (2011 – 2013)
Stroke Survivor Census Data (2014)
Census Data (2005 and 2015)
Exclusions
None
None
None
Input Data
Age-specific counts of stroke deaths for 2011-2013 period
Age-specific prevalence of stroke survivors
Age-specific total population counts for 2005 and 2015
Interpolation
Estimate population in 2012 based on age-specific change between 2005 and 2015
Output Data
Age-specific counts of stroke deaths for 2011-2013 period
Age-specific prevalence of stroke survivors in 2014
Age-specific population counts for 2012
External Data Sources
Global Burden of Disease standard life table
Global Burden of Disease average disability weight for long-term stroke survivors (0.266)
Estimate absolute YLLs, YLDs, and DALYs
Estimate YLLs: Multiply age-specific stroke deaths by remaining life expectancy in GBD standard life table
Estimate YLDs: First multiply age-specific prevalence of stroke in 2014 by 2012 population to generate expected stroke survivors in 2012.
Next, multiply count of survivors by GBD disability weight
Estimate DALYs: YLLs + YLDs
Estimate YLLs, YLDs, and DALYs per 100,000
Divide YLLs, YLDs, and DALYs by person-years lived

10. Provide a detailed description of all steps of the analysis, including mathematical formulae. This description should cover, as relevant, data cleaning, data pre-processing, data adjustments and weighting of data sources, and mathematical or statistical model(s).

Data cleaning

Data were checked for any missing variables and outliers during data cleaning phase. Missing or incorrect information was recollected from the field.

Data pre-processing

We conducted two data pre-processing actions: estimating the number of individuals alive at each age in 2012 based on the 2005 and 2012 censuses and estimating the number of individuals with stroke in 2012 based on the 2014 data.

To estimate the number of number of individuals alive in 2012, we first estimated the age-specific growth rates of the population between 2005 and 2015 using the following expression

$$r_i = \ln\left(\frac{N_i(2015)}{N_i(2005)}\right) * \frac{1}{10}$$

where r_i is the age-specific growth rate, $N_i(2015)$is the number of people in age group *i* in 2015 and $N_i(2005)$is the number of people in age group *i* in 2005. We then estimated the population in 2012 by age group as:

$$N_i(2012) = N_i(2005)e^{r_i 7}$$

Second, our estimates of stroke survivors came from a census of 45,053 individuals from 39 of the 86 villages in the surveillance site in the year 2014. To estimate the number of stroke survivors in 2012 in the overall population, we multiplied the age-specific prevalence rates from the 2014 data by the 2012 age-specific population counts.

Mathematical models

We estimated DALYs directly using the following equations:

$$YLL = \sum_{i:0,85,5} D_i * LE_i$$

where $D_i$is the number of stroke deaths in five-year age group i, and LE_i is life expectancy at age group *i* from the 2017 GBD reference life table.

$$YLD = \sum_{i:0,85,5} S_i * DW$$

where s_i are the number of stroke survivors in age group *i* and *DW* is the disability weight for stroke. After calculating both YLLs and YLDs, we calculated DALYS as

$$DALYs = YLL + YLD$$

13. Describe methods for calculating uncertainty of the estimates. State which sources of uncertainty were, and were not, accounted for in the uncertainty analysis.

Our study has three input data sources: information on counts of individuals from a population census, information on counts of deaths due to stroke from continuous population surveillance system, and information on the number of stroke survivors from a population screening of 39 of surveillance site villages. Since all three of the input sources are from census or complete population surveillance records, they do not carry with them a sampling uncertainty. This is in contrast to GBD estimates, which are based on predictions and extrapolations from statistical models rather than census data and thus carry substantial uncertainty and need to be reported with uncertainty intervals. For this reason, we do not have a measure of uncertainty around our estimate.

There are two sources of non-sampling uncertainty that we have not explicitly accounted for: uncertainty in the estimates of people alive in 2012 and uncertainty in the true number of individuals with stroke in 2012. For both these sources of uncertainty, it is not clear which direction the bias may go nor how to account for this uncertainty quantitatively. We believe that the magnitude of error introduced by these two sources of uncertainty is minor however, for the following reasons. First, our estimates of number of people alive by age group in 2012 would only produce large errors if there were sudden in or out migration events between 2005 and 2015. Second, our estimates of stroke survivors would only introduce error if the prevalence rates of stroke changed drastically between 2012 and 2014, which we believe is unlikely.

14. State how analytic or statistical source code used to generate estimates can be accessed.

We have uploaded a spreadsheet that contains generates our estimates with the submission of the manuscript.

15. Provide published estimates in a file format from which data can be efficiently extracted.

We have uploaded our tables in a spreadsheet format with the submission of the manuscript.