This document is intended to serve as a basic guide for using the Global Burden of Disease (GBD) Data Input Sources Tool. The tool lets you explore GBD 2017 input sources by GBD component, geography, and cause, risk, covariate, or impairment. After you have made your selection, you can view and access catalog entries for input sources used by GBD through the Global Health Data Exchange (GHDx).

You can download these input sources as a CSV file to see more information about how they were used in the analysis of the GBD. This CSV file contains metadata about the input sources as suggested in the [Guidelines for Accurate and Transparent Health Estimates Reporting (GATHER)](http://gather-statement.org/), a statement that promotes best practices in reporting health estimates.

# Contents

[Contents 1](#_Toc463524730)

[Accessing the tool 1](#_Toc463524731)

[Section 1: Forming a query 2](#_Toc463524732)

[Section 2: Global Health Data Exchange (GHDx) results 2](#_Toc463524733)

[Section 3: Downloading results 2](#_Toc463524734)

[Section 4: Large results 3](#_Toc463524735)

[Section 5: CSV content 3](#_Toc463524736)

[Troubleshooting 4](#_Toc463524737)

[Combining multiple CSV files 4](#_Toc463524738)

[Opening CSV files with special characters in Excel 5](#_Toc463524739)

[Produce a list of unique citations for input sources 7](#_Toc463524740)

# Accessing the tool

Access the tool through any Internet browser at [ghdx.healthdata.org/gbd-2017/data-input-sources](file:///C%3A%5CUsers%5Cjwill88%5CAppData%5CLocal%5CMicrosoft%5CWindows%5CINetCache%5CContent.Outlook%5CGCRLWDEN%5Cghdx.healthdata.org%5Cgbd-2015%5Cdata-input-sources).

# Section 1: Forming a query

Begin by selecting either a GBD component or a geography. The tool does not allow downloads for “All GBD components” or “Global.”

Options for GBD components are:

* All GBD
* Causes of Death and Illness (includes Causes of Death and Nonfatal Health Outcomes)
	+ Causes of Death
	+ Nonfatal Health Outcomes
* Covariates
* Mortality and Population
	+ Mortality
	+ Fertility
	+ Population
* Risk Factors
* Sustainable Development Goals (SDGs)

Location options are all locations in the standard GBD reporting hierarchy, including regions and some subnational locations.

Depending on the GBD component you select, you are given options to refine your query by cause, impairment, risk, or covariate. You can refine by geography at any time.

Note: If you plan to download the results of your query, it is recommended that you make the narrowest selection possible. Because the download represents the use of input data at essentially the data point level, many result sets can be millions of rows.

# Section 2: Global Health Data Exchange (GHDx) results

After you submit a query, you see an alphabetical list of the input sources used for the part of GBD you specified. Clicking on a title of any of these citations brings you the GHDx entry describing that source, including geography, time period, and topics covered by the dataset.

When possible, IHME negotiates with data owners to make the data publicly available through the GHDx. If providing the data is not possible, the record directs you to the provider of the data. For more about the GHDx, visit [About the GHDx](http://ghdx.healthdata.org/about-ghdx).

# Section 3: Downloading results

Once the list of input sources for the selection have been retrieved, you have the option to download results. Clicking on the “download” button will begin the download process for relatively small result sets.

Even for small result sets, collating results in all dimensions for a set of input sources can take some time. The download page will remain active as your downloads are prepared. When they are prepared, they will appear sequentially on the download page.

# Section 4: Large results

Queries can produce large result sets. In these cases, the tool requires an email address and your choice of how many rows you want per file. The maximum number of rows for a single file is 1 million. This number determines the number of files your download returns, not the number of total results in that download.

A large request may take up to several hours to complete. You receive a confirmation email with a link to a page where you may check the status of your request. When the process is complete, that page contains the files’ download links. You receive a second email confirming that your download is ready.

# Section 5: CSV content

While the GHDx results contain metadata on the dataset itself, the CSV generated from your query includes additional information about the way the dataset was used in producing GBD estimates. Below is a description of what the columns in the CSV represent.

|  |  |
| --- | --- |
| **Column Header** | **Description** |
| Citation | Suggested citation for the input source |
| Component | GBD Component |
| Subcomponent | Additional information on what an input was used for in analysis (e.g., severity splits, risk exposure, etc.) |
| Location | Geographic location for which the input source was used |
| Cause | Cause, etiology, or impairment for which the input source was used (where relevant) |
| Risk | Risk for which the input source was used (where relevant) |
| Covariate | Covariate for which the input source was used (where relevant) |
| Publication status | Indication if the input source is unpublished or forthcoming (if blank, the input source is considered published) |
| Indicator | SDG for which the input source was used (where relevant) |
| Provider | Provider of the data (provided whenever possible except for scientific literature) |
| Provider URL | Web address where you can find or inquire about the input source |
| GHDx URL | Web address for the GHDx record of the input source |
| Secondary GHDx URL | Web address for the GHDx record of the granular input source (where relevant)Note: In some cases, a large input source is cataloged at a granular level to increase transparency on what parts of the source were used. This is generally signified by a citation that says “as it appears in.” The secondary GHDx URL points you to catalog records for the more granular input source. For example, the WHO Mortality Database represents many country years of vital registration data, and each country year is documented by a separate GHDx catalog record. |
| Data Collection Method  | Information on the mechanism of data collection |
| Year Start | Starting year of the data derived from the input source |
| Year End | Ending year of the data derived from the input source |
| Sex  | Sex of the population of the data derived from the input source  |
| Age Start | Numerical value of starting age of the population of the data derived from the input source |
| Age End | Numerical value of ending age of the population of the data derived from the input source |
| Age Type | Unit of age values in previous columns (e.g., years, months, days) |
| Representativeness | Relevant quality of the population in the input source (e.g., nationally representative, representative of urban areas only, etc.) |
| Urbanicity Type | Urbanicity value of the population in the input source (e.g., urban, rural, etc.) |
| Population Representativeness Covariates | Other qualities of the population in the input source (where relevant and collected; see the code values and descriptions document on the tool homepage for value translation) |
| Diagnostic Criteria/Measurement Method | Relevant case definition of the population in the input source (where relevant and collected) |
| Diagnostic Criteria/Measurement Method Covariates | Case definitions of the population in the input source (where relevant and collected; see the code values and descriptions document on the tool homepage for value translation) |
| Sample Size | Sample size of data derived from input source |
| Sample Size Unit | Unit of sample size |
| Standard Error | Standard error of data derived from input source |

# Troubleshooting

This section includes workarounds for some of the known downloading issues with the GBD Data Input Source Tool.

## Combining multiple CSV files

Sometimes, queries return data in multiple CSV files. Each file contains a segment of the data and has identical columns and a header row. Note that you cannot have more than 1 million rows in a csv file. Combining results over one million rows requires use of another program such as Stata.

To combine multiple CSV files in Excel:

1. Open the first file in the set you want to combine in Excel.
2. Open the next file in the set.
3. Select all cells below the first line in this next file.

**Note**: A simple way to do this is to select cell A2 (the first and leftmost data-containing cell) and press the following keys: Ctrl + Shift + ↓ + →.

1. Copy the selected cells.
2. Paste the copied cells into the first and leftmost blank cell in the first file.
3. Repeat steps 2-5 until the data from all files has been copied into the first one.

## Opening CSV files with special characters in Excel

Sometimes, queries result in files that contain special characters, such as place names with accents like “México.” These characters may not display properly when the CSV file is opened in Excel.

To solve this problem:

1. Open Notepad.
2. Click File > Open (Ctrl + O).



1. Browse to and open the exported CSV, making sure to select "All Files" in the bottom right.



1. In the Encoding drop-down menu at the bottom, change from ANSI to UTF-8.



1. Save the file. When you reopen it in Excel, the special characters should render correctly.

## Produce a list of unique citations for input sources

In order to provide metadata on how input sources were used in GBD analysis, a single input source is represented by multiple rows in a CSV download. To get a simple list of unique citations for input sources without the additional metadata, you can remove duplicates in your CSV based on the citation column.



Depending on the number of CSVs returned by your query, you may need to do this for multiple sheets, then combine the content of all CSVs into a single CSV and remove duplicates a final time.