Data Profile - World Health Organization API

Kurt Wydrinski

11/22/2019

Source

During the formation of the United Nations in 1945, one of the key objectives for its ongoing mission was to foster cooperation between countries to address many social well-being issues, including health.[1] This objective was taken on as part of the World Health Organization (WHO) which was established on April 7, 1948.[2] The WHO continues this mission today with 194 countries cooperating at various levels to promote global health concerns. In doing so, the WHO regularly collects data from its members that are related to health issues. The data contains various time series regarding diseases, illnesses, economics, social demographics, etc. This data is collectively captured in the WHO's online database *Global Health Observatory (GHO)*.[3]

API Overview

The R package `WHO`, provides a simple API to access the GHO. It only provides two functions: `get_codes()` and `get_data()`. Inside the GHO, each time series that exists is identifed by a label. Each label is a code that uniquely identifies the series. These labels are then used a a parameter to `get_data()` to retrieve the time series observations.

```
# install.packages("WHO")
library(WHO)
```

The code below uses the `extra` parameter to download all metadata available for the GHO codes.

```
who_codes <- get_codes(extra = TRUE)
glimpse(who_codes)
```

```
## Observations: 3,287
## Variables: 9
## $ label          <chr> "MDG_0000000001", "MDG_0000000003", "MDG_0000000005"...
## $ display        <chr> "Infant mortality rate (probability of dying between...
## $ url            <chr> "https://www.who.int/data/gho/indicator-metadata-reg...
## $ display_fr     <chr> "Taux de mortalité des nourrissons (probabilité de d...
## $ display_es     <chr> "Tasa de mortalidad de menores de 1 año (probabilida...
## $ definition_xml <chr> "http://apps.who.int/gho/indicatorregistryservice/pu...
## $ category       <chr> "Mortality and global health estimates", "Sustainabl...
## $ imr_id         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "16", NA...
## $ renderer_id    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
```

There are 3,287 indicators or time series in this dataset. The `category` variable is a WHO grouping of the indicators into 51 sets of indicators as follows:

```
who_codes %>%
  group_by(category) %>%
  summarise(indicator_count = n()) %>%
  knitr::kable()
```

[1] See https://www.history.com/this-day-in-history/the-united-nations-is-born.
[2] See https://www.who.int/about/who-we-are/history.
[3] See https://www.who.int/gho/about/en/.

category	indicator_count
AMR GLASS Coordination	4
AMR GLASS Quality assurance	5
AMR GLASS Surveillance	3
Child health	37
Demographic and socioeconomic statistics	19
Essential health technologies	10
FINANCIAL PROTECTION	28
Global Observatory for eHealth (GOe)	92
Health Equity Monitor	64
Health financing	13
Health systems	381
Health workforce	24
HIV/AIDS and other STIs	37
Infectious diseases	31
Infrastructure	6
Injuries and violence	59
Insecticide resistance	9
International Health Regulations (2005) monitoring framework	13
Malaria	126
Medical equipment	8
Millennium Development Goals (MDGs)	15
Mortality and global health estimates	58
Negelected tropical diseases	9
Neglected tropical diseases	17
Neglected Tropical Diseases	17
Noncommunicable diseases	50
Noncommunicable diseases and mental health	2
Noncommunicable diseases CCS	82
Nutrition	27
Oral health	1
Public health and environment	177
RSUD: GOVERNANCE, POLICY AND FINANCING : PREVENTION	3
RSUD: GOVERNANCE, POLICY AND FINANCING: FINANCING	7
RSUD: GOVERNANCE, POLICY AND FINANCING: TREATMENT	12
RSUD: HUMAN RESOURCES	8
RSUD: INFORMATION SYSTEMS	6
RSUD: SERVICE ORGANIZATION AND DELIVERY: PHARMACOLOGICAL TREATMENT	11
RSUD: SERVICE ORGANIZATION AND DELIVERY: PREVENTION PROGRAMS AND PROVIDERS	5
RSUD: SERVICE ORGANIZATION AND DELIVERY: SCREENING AND BRIEF INTERVENTIONS	4
RSUD: SERVICE ORGANIZATION AND DELIVERY: SPECIAL PROGRAMMES AND SERVICES	12
RSUD: SERVICE ORGANIZATION AND DELIVERY: TREATMENT CAPACITY AND TREATMENT COVERAGE	2
RSUD: SERVICE ORGANIZATION AND DELIVERY: TREATMENT SECTORS AND PROVIDERS	8
RSUD: YOUTH	3
Substance use and mental health	1121
Sustainable development goals	28
Tobacco	3
Tuberculosis	68
Universal Health Coverage	16
Urban health	32
World Health Statistics	67
NA	447

Note the following about the above categories.

1. There are 447 indicators that do not have a category assigned by the WHO. These should be considered as a category called `Uncategorized` or `No category`.
2. The naming convention is not consistent across categories. For example, some use single words, phrases, all capital letters, irregular case, etc. Category names should be somewhat standardized while retaining the original meaning in the dataset.
3. A number of categories appear to be fragmented. For example, there are three premutations of *Neglected Tropical Diseases* that need to be combined into a single category.

The table below provides a list of the indicators in the *Demographic and socioeconomic statistics* category. However, inspection of the table should make it apparent that a number of these indicators are based on population.

```
who_codes %>%
  filter(category == "Demographic and socioeconomic statistics") %>%
  select(label, display) %>%
  knitr::kable()
```

label	display
WHS9_CBR	Crude birth rate (per 1000 population)
WHS9_CDR	Crude death rate (per 1000 population)
WHS9_CS	Cellular subscribers (per 100 population)
WHS10_1	Most recent census (year)
WHS10_2	Number of cause-of-death registration years available
WHS10_3	Number of national population surveys - child anthropometry
WHS10_4	Number of national population surveys - child mortality

label	display
WHS10_5	Number of national population surveys - maternal mortality
WHS10_6	Number of national population surveys - HIV prevalence
WHS10_7	Number of national population surveys - adult health
WHS10_8	Civil registration coverage of cause-of-death (%)
WHS10_9	Ill-defined causes in cause-of-death registration (%)
CCO_1	Poverty headcount ratio at $1.25 a day (PPP) (% of population)
CCO_2	Human development index rank
CCO_3	Gender inequality index rank
ITU_IDI	ICT Development Index (IDI)
ITU_IDI_RANK	ICT Development Index (IDI) rank
ITU_ICT_1	Percentage of individuals using the Internet
ITU_ICT_2	Mobile-cellular telephone subscriptions per 100 inhabitants

A closer look may make it apparent that the data may not have a total population number. It is possible that `WHS10_1` could have it by the `display` string.

```
who_codes %>%
  filter(label == "WHS10_1")
```

```
## # A tibble: 1 x 9
##   label display url   display_fr display_es definition_xml category imr_id
##   <chr> <chr>   <chr> <chr>      <chr>      <chr>          <chr>    <chr>
## 1 WHS1~ Most r~ http~ <NA>       <NA>       http://apps.w~ Demogra~ <NA>
## # ... with 1 more variable: renderer_id <chr>
```

Inspection of the indicator observation in the metadata shows that the `url` and `definition_xml` variables are URLs to further information about the indicator. `url` is an address for a web page on the GHO registry that explains what the indicator represents. The `defintion_xml` provides the same information presented in the explanation page in XML format.

As the explanation page suggests, the `WHS10_1` indicator does reflect the population count/census statistic. This may or may not be useful when used with other indicators. In looking at all indicators in the metadata, there are at least three others that are called `Population`.

```
who_codes %>%
  filter(str_detect(display, "^Population$")) %>%
  select(label, display, category)
```

```
## # A tibble: 3 x 3
##   label         display    category
##   <chr>         <chr>      <chr>
## 1 MALARIA_15279 Population Malaria
## 2 RS_1845       Population Injuries and violence
## 3 MEDS1_01_01   Population Health systems
```

API Risks

Based on the analysis above, the following risks appear to exist when using this API dataset.

1. The data is not tidy and needs to be made such before detailed analysis can be completed.
2. The structure of the data is inconsistent and needs detailed exploration when deciding on which indicators to use.

3. The `category` variable in the metadata (i.e. codes) is not very reliable. Careful considerations need to be made when leveraging this variable. The category should be used as the basis for a new variable which is a reliable identifier for logically-related indicators.

API Rewards

Based on the analysis above the following benefits seem to be gained by using this API and dataset.

1. The data can be accessed using the simple `WHO` package.
2. The data is acessible via the Internet via on-demand API calls.
3. The data contains international health and disease data along with related indicators.

Usage Prototype

The sections below provide a prototypical usage for this dataset and API. The actual use and visualization of the data will vary between projects. However, the sections below demonstrate common usage patterns for this data and API.

Prototype Exploration

per 100,000 Indicators

A number of the indicators in the database are measured and scaled to units of 100,000. Most these indicators are scaling to report over country population units of 100,000 people. The table below presents a sampling of these indicators.

```
who_codes %>%
  filter(str_detect(display, "per 100 000 ")) %>%
  select(label, display, category)
```

```
## # A tibble: 90 x 3
##    label      display                                      category
##    <chr>      <chr>                                        <chr>
##  1 MDG_00000~ Deaths due to malaria (per 100 000 populat~ Millennium Developmen~
##  2 MDG_00000~ Deaths due to tuberculosis among HIV-negat~ Tuberculosis
##  3 MDG_00000~ Deaths due to tuberculosis among HIV-posit~ Millennium Developmen~
##  4 MDG_00000~ Prevalence of HIV among adults aged >=15 y~ Millennium Developmen~
##  5 MDG_00000~ Incidence of tuberculosis (per 100 000 pop~ Tuberculosis
##  6 MDG_00000~ Prevalence of tuberculosis (per 100 000 po~ Tuberculosis
##  7 MDG_00000~ Maternal mortality ratio (per 100 000 live~ Mortality and global ~
##  8 MORT_61    Mortality - crude death rate per 100 000 p~ Mortality and global ~
##  9 WHS2_131   Age-standardized NCD mortality rate (per 1~ Mortality and global ~
## 10 WHS2_138   Deaths due to HIV/AIDS (per 100 000 popula~ World Health Statisti~
## # ... with 80 more rows
```

Usefulness Assessment

There are about ninety indicators that report statistics on groups of 100,000. Most of the grouping are by population (i.e. *per 100,000 people*) but there are also a few by *adult population* and *live births*. As can be seen in the table below, these indicators cross a number of topics. It could be possible to tell a story about *The X per 100,000 people...*.

Topic	Indicators
HIV/AIDS and other STIs	2
Infectious diseases	5
Infrastructure	6
Injuries and violence	2
Malaria	1
Millennium Development Goals (MDGs)	4
Mortality and global health estimates	13
Neglected Tropical Diseases	1
Public health and environment	14
Substance use and mental health	19
Sustainable development goals	8
Tuberculosis	6
World Health Statistics	7
NA	2

Infectious diseases Indicators

A number of the indicators in the database measure various statistics about different infectious diseases. There are statistics for number of cases, number of deaths from various infectious diseases, and a few other more detailed statsitics about diseases such as leptospirosis (not to be confused with leprosy.) The table below presents a sampling of these indicators.

```
who_codes %>%
  filter(category == "Infectious diseases")
```

```
## # A tibble: 31 x 9
##    label display url   display_fr display_es definition_xml category imr_id
##    <chr> <chr>   <chr> <chr>      <chr>      <chr>          <chr>    <chr>
##  1 WHS3~ Choler~ http~ <NA>       <NA>       http://apps.w~ Infecti~ <NA>
##  2 WHS3~ Diphth~ http~ <NA>       <NA>       http://apps.w~ Infecti~ <NA>
##  3 WHS3~ Japane~ http~ <NA>       <NA>       http://apps.w~ Infecti~ <NA>
##  4 WHS3~ Pertus~ http~ <NA>       <NA>       http://apps.w~ Infecti~ <NA>
##  5 WHS3~ Number~ http~ <NA>       <NA>       http://apps.w~ Infecti~ <NA>
##  6 WHS3~ Total ~ http~ <NA>       <NA>       http://apps.w~ Infecti~ <NA>
##  7 WHS3~ Mening~ http~ <NA>       <NA>       http://apps.w~ Infecti~ <NA>
##  8 WHS3~ Poliom~ http~ <NA>       <NA>       http://apps.w~ Infecti~ <NA>
##  9 WHS3~ Yellow~ http~ <NA>       <NA>       http://apps.w~ Infecti~ <NA>
## 10 WHS3~ H5N1 i~ ""    <NA>       <NA>       http://apps.w~ Infecti~ 53
## # ... with 21 more rows, and 1 more variable: renderer_id <chr>
```

Each of these indicators has a number of time series with annual values for different countries. However, not all of these indicators have a lot of data and may not be very usable. The sections below describe a few indicators that are of interest and could be of some use.

"The Bird Flu" - Number of Cases

The `WHS3_51` indicator shows the *H5N1 influenza - number of reported cases*. The H5N1 strain of influenza is commonly know as *avian flu* or *the bird flu*.

```
# This function is used to create clean factors for analysis from the GHO
# dataset.
gho_vectorize <- function(v) {
  switch(class(v),
```

```
        "character" = {
          f <- as_factor(v)
          fct_explicit_na(fct_relevel(f,
                                      sort(levels(f))),
                          na_level = "NA")
        },
        "factor" = {
          fct_explicit_na(fct_relevel(v,
                                      sort(levels(v))),
                          na_level = "NA")
        },
        "numeric" = as_factor(v))
}
```

```
tb_bird_flu <- get_data("WHS3_51")
tb_bird_flu$country <- gho_vectorize(tb_bird_flu$country)
tb_bird_flu$region <- gho_vectorize(tb_bird_flu$region)
tb_bird_flu$publishstate <- gho_vectorize(tb_bird_flu$publishstate)
```

```
tb_bird_flu %>%
  group_by(country) %>%
  select(country, year, value, worldbankincomegroup, region) %>%
  arrange(country, year)
```

```
## # A tibble: 28 x 5
## # Groups:   country [16]
##    country     year value worldbankincomegroup region
##    <fct>      <dbl> <dbl> <chr>                <fct>
##  1 Azerbaijan  2012     0 <NA>                 Europe
##  2 Bangladesh  2011     2 <NA>                 South-East Asia
##  3 Bangladesh  2012     3 <NA>                 South-East Asia
##  4 Cambodia    2010     1 Low-income           Western Pacific
##  5 Cambodia    2011     8 <NA>                 Western Pacific
##  6 Cambodia    2012     3 <NA>                 Western Pacific
##  7 China       2010     2 Lower-middle-income  Western Pacific
##  8 China       2011     1 <NA>                 Western Pacific
##  9 China       2012     2 <NA>                 Western Pacific
## 10 Djibouti    2012     0 <NA>                 Eastern Mediterranean
## # ... with 18 more rows
```

```
summary(tb_bird_flu)
```

```
##      gho                 year           country                      region
##  Length:28          Min.   :2010   Cambodia  : 3   Africa               : 1
##  Class :character   1st Qu.:2011   China     : 3   Eastern Mediterranean: 6
##  Mode  :character   Median :2012   Egypt     : 3   Europe               : 2
##                     Mean   :2011   Indonesia : 3   South-East Asia      : 7
##                     3rd Qu.:2012   NA        : 3   Upper-middle-income  : 1
##                     Max.   :2012   Bangladesh: 2   Western Pacific      :11
##                                    (Other)   :11
##     publishstate     value        worldbankincomegroup
##  Published:28    Min.   : 0.000   Length:28
##                  1st Qu.: 0.000   Class :character
##                  Median : 2.000   Mode  :character
##                  Mean   : 5.821
```

```
##                  3rd Qu.: 9.000
##                  Max.   :39.000
##
```

Usefulness Assessment

The following items may cause issues when using this indicator.

1. There is are only 28 observations for Bird Flu in the dataset.
2. There are 3 observations with `NA` for the country. 2.1. Two of these are for the *Western Pacific* region which also includes Viet Nam, China, etc. These oberservations could be reports of incidents that could not be attributed to a given country in the region but may have been reported by one. For example, a foreign traveller could have been treated for the incident in Viet Nam. It is not fair to assume the person contracted the disease in Viet Nam so the `NA` country should be considered a valid observation. 2.2. The third of these is attributed to the *Upper-middle-income* WHO income group category. No other observations have this attribution so this one appears to be an observation caputured by the WHO to highlight the group.

Based on the issues above, there is some elevated risk in using this dataset. It is very small so statistics will be hard to accurately derive. The `NA` country observations suggest a mixture of hetergeneous observations; most are observations of the number of cases in a given country for a given year while others summarize case reports for various groups that span countries or outside of countries in some way.

Tuberculosis - Number of Cases

The `WHS3_54` indicator shows the *Number of reported cases of tuberculosis (DOTS).*

```
tb_tuber <- get_data("WHS3_54")
tb_tuber$country <- gho_vectorize(tb_tuber$country)
tb_tuber$region <- gho_vectorize(tb_tuber$region)
tb_tuber$publishstate <- gho_vectorize(tb_tuber$publishstate)
```

```
tb_tuber %>%
  group_by(country) %>%
  arrange(country, year) %>%
  select(country, year, value, worldbankincomegroup, region)
```

```
## # A tibble: 188 x 5
## # Groups:   country [188]
##    country              year value worldbankincomegroup region
##    <fct>               <dbl> <dbl> <chr>                <fct>
##  1 Afghanistan          2008 13136 <NA>                 Eastern Mediterranean
##  2 Albania              2008   170 <NA>                 Europe
##  3 Algeria              2008  8643 <NA>                 Africa
##  4 Andorra              2008     3 <NA>                 Europe
##  5 Angola               2008 22562 <NA>                 Africa
##  6 Antigua and Barbuda  2008     1 <NA>                 Americas
##  7 Argentina            2008  4758 <NA>                 Americas
##  8 Armenia              2008   487 <NA>                 Europe
##  9 Australia            2008   299 <NA>                 Western Pacific
## 10 Azerbaijan           2008  1409 <NA>                 Europe
## # ... with 178 more rows
```

```
summary(tb_tuber)
```

```
##      gho                 year                     country
```

```
##  Length:188         Min.   :2008   Afghanistan        :  1
##  Class :character   1st Qu.:2008   Albania            :  1
##  Mode  :character   Median :2008   Algeria            :  1
##                     Mean   :2008   Andorra            :  1
##                     3rd Qu.:2008   Angola             :  1
##                     Max.   :2008   Antigua and Barbuda:  1
##                                    (Other)            :182
##                   region      publishstate     value
##  Africa               :46   Published:188   Min.   :     0.0
##  Americas             :35                   1st Qu.:   146.2
##  Eastern Mediterranean:21                   Median :  1338.0
##  Europe               :50                   Mean   : 14119.2
##  South-East Asia      :11                   3rd Qu.:  6274.8
##  Western Pacific      :25                   Max.   :615977.0
##
##  worldbankincomegroup
##  Length:188
##  Class :character
##  Mode  :character
##
##
##
##
```

```
who_codes %>%
  filter(label == "WHS3_54")
```

```
## # A tibble: 1 x 9
##   label display url   display_fr display_es definition_xml category imr_id
##   <chr> <chr>   <chr> <chr>      <chr>      <chr>          <chr>    <chr>
## 1 WHS3~ Number~ http~ <NA>       <NA>       http://apps.w~ Infecti~ <NA>
## # ... with 1 more variable: renderer_id <chr>
```

Usefulness Assessment

The following item may cause issues when using this indicator.

1. There are only observations for the year 2008.

Based on the issue above, this data will have limited use bacause it only covers one year.

Cholera - Number of Cases

The `CHOLERA_0000000001` indicator shows the *Number of reported cases of cholera*.

```
tb_cholera <- get_data("CHOLERA_0000000001")
tb_cholera$country <- gho_vectorize(tb_cholera$country)
tb_cholera$region <- gho_vectorize(tb_cholera$region)
tb_cholera$publishstate <- gho_vectorize(tb_cholera$publishstate)
```

```
tb_cholera %>%
  group_by(country) %>%
  arrange(country, year) %>%
  select(country, year, value, worldbankincomegroup, region)
```

```
## # A tibble: 2,480 x 5
```

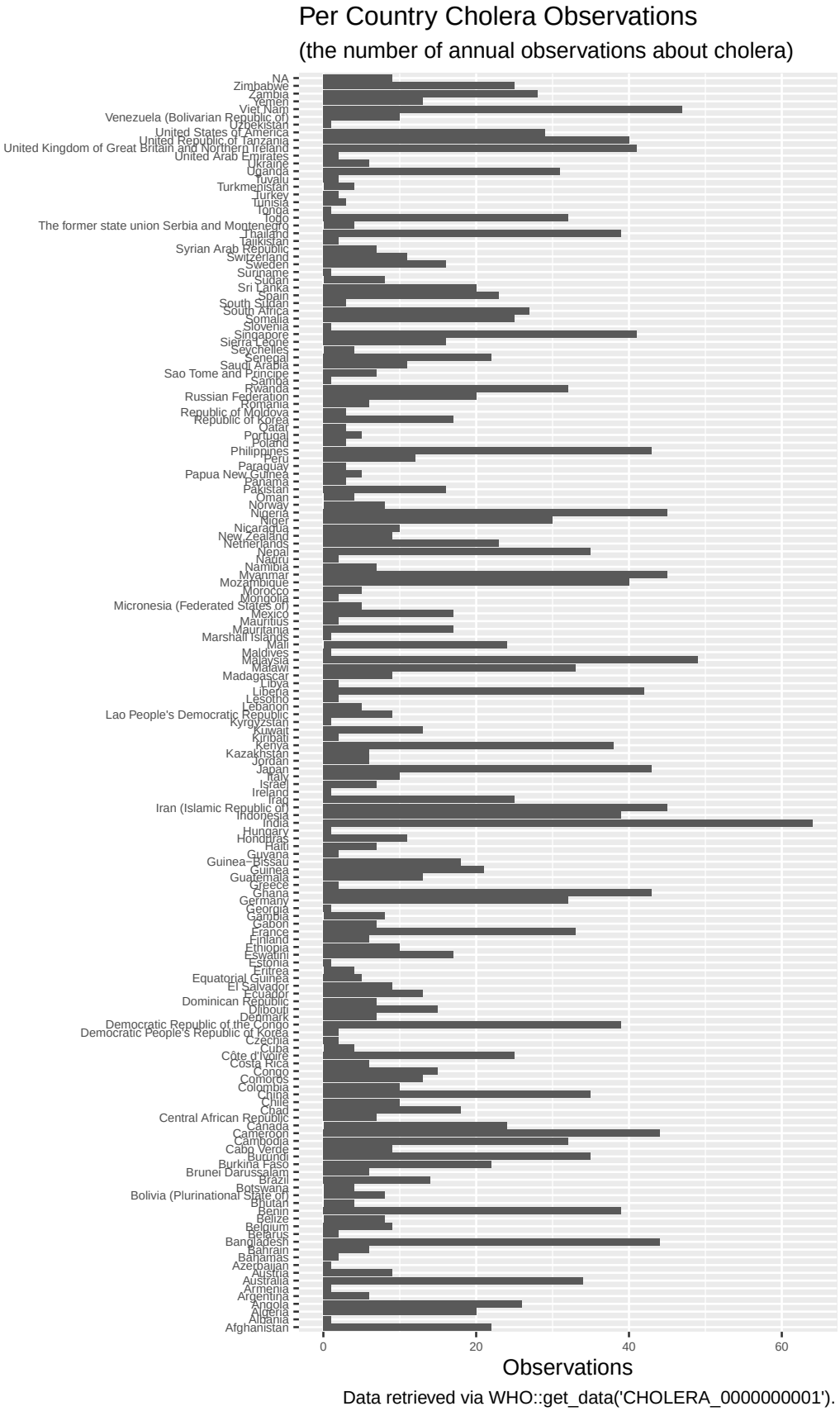
```
## # Groups:   country [162]
##    country      year value worldbankincomegroup region
##    <fct>       <dbl> <dbl> <chr>                <fct>
##  1 Afghanistan  1960   887 <NA>                 Eastern Mediterranean
##  2 Afghanistan  1965   218 <NA>                 Eastern Mediterranean
##  3 Afghanistan  1993 37046 <NA>                 Eastern Mediterranean
##  4 Afghanistan  1994 38735 <NA>                 Eastern Mediterranean
##  5 Afghanistan  1995 19903 <NA>                 Eastern Mediterranean
##  6 Afghanistan  1997  4170 <NA>                 Eastern Mediterranean
##  7 Afghanistan  1998 10000 <NA>                 Eastern Mediterranean
##  8 Afghanistan  1999 24639 <NA>                 Eastern Mediterranean
##  9 Afghanistan  2000  4330 <NA>                 Eastern Mediterranean
## 10 Afghanistan  2001  4499 <NA>                 Eastern Mediterranean
## # ... with 2,470 more rows
```

```
summary(tb_cholera)
```

```
##      gho                 year                              country
##  Length:2480        Min.   :1949   India                        :  64
##  Class :character   1st Qu.:1981   Malaysia                     :  49
##  Mode  :character   Median :1995   Viet Nam                     :  47
##                     Mean   :1993   Iran (Islamic Republic of):  45
##                     3rd Qu.:2005   Myanmar                      :  45
##                     Max.   :2016   Nigeria                      :  45
##                                    (Other)                      :2185
##                    region        publishstate      value
##  Africa               :1005   Published:2480   Min.   :     0
##  Western Pacific      : 388                    1st Qu.:    10
##  Europe               : 312                    Median :   244
##  South-East Asia      : 294                    Mean   :  3816
##  Americas             : 240                    3rd Qu.:  1934
##  Eastern Mediterranean: 237                    Max.   :340311
##  (Other)              :   4
##  worldbankincomegroup
##  Length:2480
##  Class :character
##  Mode  :character
##
##
##
##
```

```
ggplot(data = tb_cholera %>%
                group_by(country) %>%
                summarise(count = n())) +
  geom_col(mapping = aes(x = country,
                         y = count)) +
  coord_flip() +
  theme(axis.text = element_text(size = 6)) +
  labs(title = "Per Country Cholera Observations",
       subtitle = "(the number of annual observations about cholera)",
       caption = "Data retrieved via WHO::get_data('CHOLERA_0000000001').",
       tag = "GHO",
       x = "",
       y = "Observations")
```

GHO


Per Country Cholera Observations
(the number of annual observations about cholera)
NA
Zimbabwe
Zambia
Yemen
Viet Nam
Venezuela (Bolivarian Republic of)
Uzbekistan
United States of America
United Republic of Tanzania
United Kingdom of Great Britain and Northern Ireland
United Arab Emirates
Ukraine
Uganda
Tuvalu
Turkmenistan
Turkey
Tunisia
Tonga
Togo
The former state union Serbia and Montenegro
Thailand
Tajikistan
Syrian Arab Republic
Switzerland
Sweden
Suriname
Sudan
Sri Lanka
Spain
South Sudan
South Africa
Somalia
Slovenia
Singapore
Sierra Leone
Seychelles
Senegal
Saudi Arabia
Sao Tome and Principe
Samoa
Rwanda
Russian Federation
Romania
Republic of Moldova
Republic of Korea
Qatar
Portugal
Poland
Philippines
Peru
Paraguay
Papua New Guinea
Panama
Pakistan
Oman
Norway
Nigeria
Niger
Nicaragua
New Zealand
Netherlands
Nepal
Nauru
Namibia
Myanmar
Mozambique
Morocco
Mongolia
Micronesia (Federated States of)
Mexico
Mauritius
Mauritania
Marshall Islands
Mali
Maldives
Malaysia
Malawi
Madagascar
Libya
Liberia
Lesotho
Lebanon
Lao People's Democratic Republic
Kyrgyzstan
Kuwait
Kiribati
Kenya
Kazakhstan
Jordan
Japan
Italy
Israel
Ireland
Iraq
Iran (Islamic Republic of)
Indonesia
India
Hungary
Honduras
Haiti
Guyana
Guinea-Bissau
Guinea
Guatemala
Greece
Ghana
Germany
Georgia
Gambia
Gabon
France
Finland
Ethiopia
Eswatini
Estonia
Eritrea
Equatorial Guinea
El Salvador
Ecuador
Dominican Republic
Djibouti
Denmark
Democratic Republic of the Congo
Democratic People's Republic of Korea
Czechia
Cuba
Côte d'Ivoire
Costa Rica
Congo
Comoros
Colombia
China
Chile
Chad
Central African Republic
Canada
Cameroon
Cambodia
Cabo Verde
Burundi
Burkina Faso
Brunei Darussalam
Brazil
Botswana
Bolivia (Plurinational State of)
Bhutan
Benin
Belize
Belgium
Belarus
Bangladesh
Bahrain
Bahamas
Azerbaijan
Austria
Australia
Armenia
Argentina
Angola
Algeria
Albania
Afghanistan
0
20
40
60
Observations
Data retrieved via WHO::get_data('CHOLERA_0000000001').

```
tb_cholera %>%
  filter(country == "NA")
```

```
## # A tibble: 9 x 7
##   gho               year country region     publishstate  value worldbankincomeg~
##   <chr>            <dbl> <fct>   <fct>      <fct>         <dbl> <chr>
## 1 Number of repo~  2013 NA      NA         Published        41 High-income
## 2 Number of repo~  2013 NA      Western P~ Published       246 <NA>
## 3 Number of repo~  2013 NA      South-Eas~ Published      6049 <NA>
## 4 Number of repo~  2013 NA      NA         Published      9474 Upper-middle-inc~
## 5 Number of repo~  2013 NA      Eastern M~ Published     12147 <NA>
## 6 Number of repo~  2013 NA      NA         Published     15442 Lower-middle-inc~
## 7 Number of repo~  2013 NA      Africa     Published     49465 <NA>
## 8 Number of repo~  2013 NA      Americas   Published     61152 <NA>
## 9 Number of repo~  2013 NA      (WHO) Glo~ Published    129067 Global
```

Usefulness Assessment

The following items may cause issues when using this indicator.

1. Not all countries have a cosnsistent number of annual observations.
2. There are 9 observations where `country` is `NA`. 2.1. These were all from 2013. 2.2. They appear to be summmary data produced in 2013 for different regions (including a "global" region).

Although the above issues pose some elevated risk in using this dataset, there are enough observations across countries and years to analyze and report on this data.

water Indicators

Cholera is caused by the *Vibrio cholerae* bacterium which has infected either food or water that has been ingested.[4] The *GHO* has a number of indicators in its datasets that could be useful to help tell a story about cholera. The table below presents a sampling of these indicators.

```
who_codes %>%
  filter(str_detect(display, "water")) %>%
  select(category, display, label) %>%
  arrange(category, label)
```

```
## # A tibble: 21 x 3
##    category              display                                          label
##    <chr>                 <chr>                                            <chr>
##  1 Public health and en~ Population using improved drinking-water sourc~ EQ_WAT~
##  2 Public health and en~ Population using improved drinking-water sourc~ WHS5_1~
##  3 Public health and en~ Number of diarrhoea deaths from inadequate wat~ WSH_10
##  4 Public health and en~ Number of diarrhoea deaths from inadequate wat~ WSH_10~
##  5 Public health and en~ Attributable fraction of diarrhoea to inadequa~ WSH_20
##  6 Public health and en~ Attributable fraction of diarrhoea to inadequa~ WSH_20~
##  7 Public health and en~ Number of diarrhoea DALYs from inadequate wate~ WSH_30
##  8 Public health and en~ Number of diarrhoea DALYs from inadequate water WSH_30~
##  9 Public health and en~ Diarrhoea deaths from inadequate water, sanita~ WSH_40
## 10 Public health and en~ Diarrhoea deaths from inadequate water in chil~ WSH_40~
## # ... with 11 more rows
```

[4] See https://www.webmd.com/a-to-z-guides/cholera-faq#1.

Water - Improved Drinking Sources (WHS5_122)

The WHS5_122 indicator shows the *Population using improved drinking-water sources (%)*.

```
tb_water_improved <- get_data("WHS5_122")
```

Usefulness Assessment

The following items will cause issues when using this indicator.

1. There are no observations available for this indicator.

Without any data, there is nothing to analyze or report. This indicator is no useful.

Water - Improved Drinking Sources (EQ_WATER)

The EQ_WATER indicator shows the *Population using improved drinking-water sources (%)*.

```
tb_water_improved <- get_data("EQ_WATER")
tb_water_improved$country <- gho_vectorize(tb_water_improved$country)
tb_water_improved$region <- gho_vectorize(tb_water_improved$region)
tb_water_improved$publishstate <- gho_vectorize(tb_water_improved$publishstate)
```

```
tb_water_improved %>%
  arrange(country, year) %>%
  select(country, year, value, wealthquintile, region)
```

```
## # A tibble: 300 x 5
##    country     year value wealthquintile region
##    <fct>      <dbl> <dbl> <chr>          <fct>
##  1 Bangladesh  1993  98.7 Q1 (Poorest)   South-East Asia
##  2 Bangladesh  1993  95.8 Q2             South-East Asia
##  3 Bangladesh  1993  99   Q3             South-East Asia
##  4 Bangladesh  1993  99.4 Q4             South-East Asia
##  5 Bangladesh  1993 100   Q5 (Richest)   South-East Asia
##  6 Bangladesh  1993  91.7 <NA>           South-East Asia
##  7 Bangladesh  2007  98.9 Q1 (Poorest)   South-East Asia
##  8 Bangladesh  2007  98.9 Q2             South-East Asia
##  9 Bangladesh  2007  99.5 Q3             South-East Asia
## 10 Bangladesh  2007  99.7 Q4             South-East Asia
## # ... with 290 more rows
```

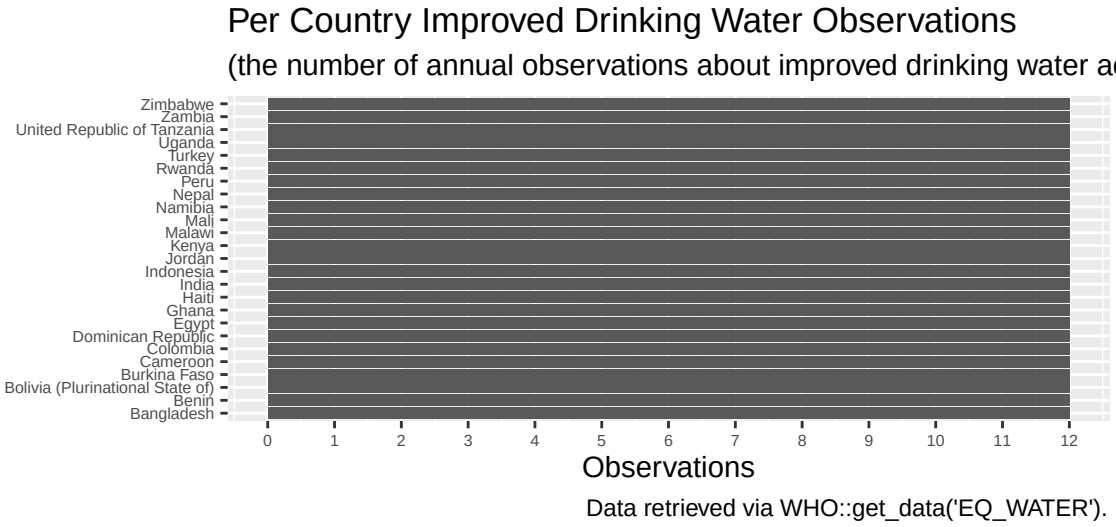
```
summary(tb_water_improved)
```

```
##      gho                 year                                   country
##  Length:300         Min.   :1990   Bangladesh                       : 12
##  Class :character   1st Qu.:1994   Benin                            : 12
##  Mode  :character   Median :2000   Bolivia (Plurinational State of): 12
##                     Mean   :1999   Burkina Faso                     : 12
##                     3rd Qu.:2005   Cameroon                         : 12
##                     Max.   :2007   Colombia                         : 12
##                                    (Other)                          :228
##  wealthquintile                        region       publishstate
##  Length:300         Africa                :156   Published:300
##  Class :character   Americas              : 60
##  Mode  :character   Eastern Mediterranean: 24
##                     Europe                : 12
```

```
##                     South-East Asia      : 48
##
##
##      value        residenceareatype
##  Min.   :  3.40   Length:300
##  1st Qu.: 57.52   Class :character
##  Median : 80.05   Mode  :character
##  Mean   : 72.97
##  3rd Qu.: 95.25
##  Max.   :100.00
##
```

```
ggplot(data = tb_water_improved %>%
                group_by(country) %>%
                summarise(count = n())) +
  geom_col(mapping = aes(x = country,
                         y = count)) +
  scale_y_continuous(breaks = seq(0, 15, 1)) +
  coord_flip() +
  theme(axis.text = element_text(size = 6)) +
  labs(title = "Per Country Improved Drinking Water Observations",
       subtitle = "(the number of annual observations about improved drinking water access)",
       caption = "Data retrieved via WHO::get_data('EQ_WATER').",
       tag = "GHO",
       x = "",
       y = "Observations")
```

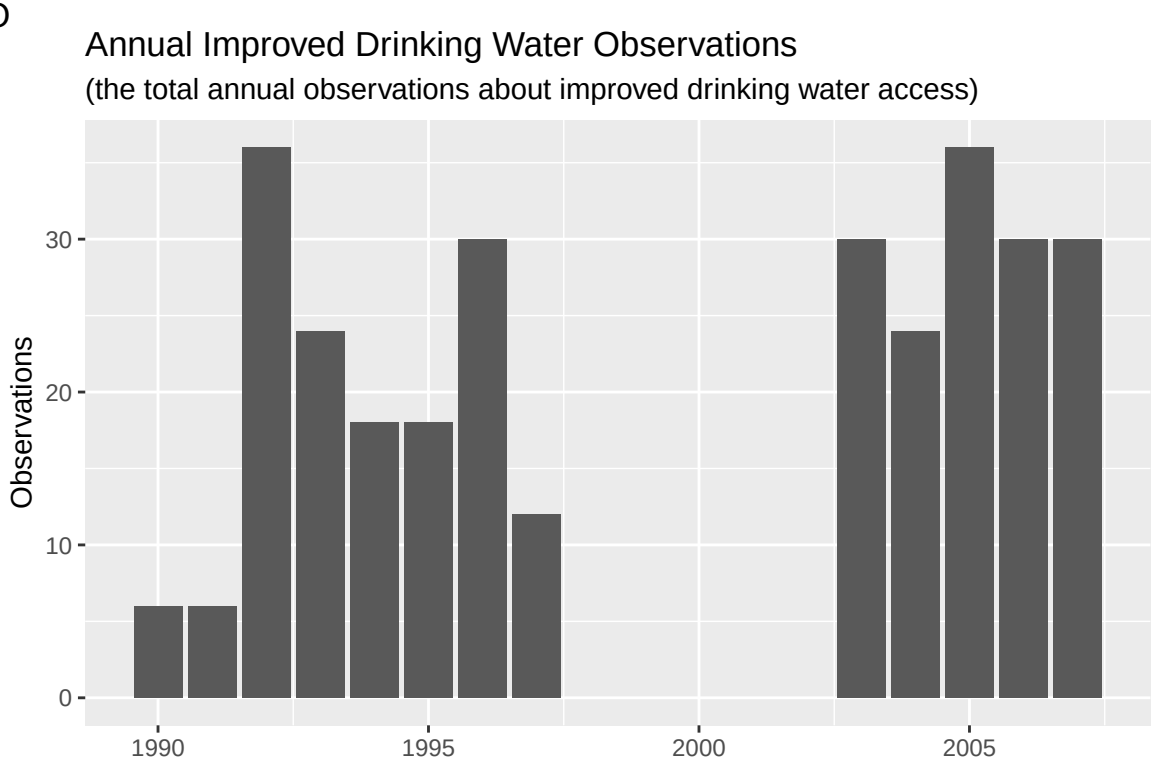
GHO




```
ggplot(data = tb_water_improved %>%
                group_by(year) %>%
                summarise(count = n())) +
  geom_col(mapping = aes(x = year,
                         y = count)) +
  labs(title = "Annual Improved Drinking Water Observations",
       subtitle = "(the total annual observations about improved drinking water access)",
       caption = "Data retrieved via WHO::get_data('EQ_WATER').",
       tag = "GHO",
       x = "",
```

```
       y = "Observations")
```

GHO



Usefulness Assessment

The following items may cause issues when using this indicator.

1. There are no observations for the years 1998-2002.
2. The dataset seems like it could be incomplete data. The WHO defintion of this indicator can be located at the WHO web site. At the bottom of the page, there is a link to a joint project between WHO and UNICEF regarding safe drinking water. This web site contains data for more countries and for more years regarding safe drinking water. It appears the `WHO` package API may not be pulling all available data.
3. The data is not tidy because each year has 5 rows with each row represent the value for a certain quintile of the wealth of the population.

Based on the issue above, this data may have limited use bacause it only covers a subset of the available data. The data at the joint project between the WHO and UNICEF may be better to use than this dataset.

Prototype Requirements

Although there are some limitations to this data, there should be enough usable data to create some interested visualizations. If this data was used in conjunction with the WHO/UNICEF data, we could tell a story with the following visualizations.

1. Overview of different infectious diseases around the world.
2. Highlight of cholera incidents and deaths around the world.
3. Breakdown of water access by wealth quintiles in cholera-affected nations.

4. TBD more detail regarding wealth (or aid) with data TBD.

We'd need further analysis but should be able to easily find data to further breakdown the wealth aspect. There many other indicator in the WHO data to analyze and possibel use for this. If they aren't usuable, we can get economic data regarding wealth elsewhere.