



Max-Planck-Institut für demografische Forschung
Max Planck Institute for Demographic Research
Konrad-Zuse-Strasse 1 · D-18057 Rostock · GERMANY
Tel +49 (0) 3 81 20 81 - 0; Fax +49 (0) 3 81 20 81 - 202;
<http://www.demogr.mpg.de>

MPIDR TECHNICAL REPORT 2010-004
MAY 2010

A VBA program for data extraction from the WHO Mortality Database

Evgeny M. Andreev (andreev@demogr.mpg.de)

This technical report has been approved for release by: Vladimir Shkolnikov (shkolnikov@demogr.mpg.de),
Head of the Laboratory of Demographic Data.

© Copyright is held by the authors.

Technical reports of the Max Planck Institute for Demographic Research receive only limited review.
Views or opinions expressed in technical reports are attributable to the authors and do not necessarily
reflect those of the Institute.

A VBA program for data extraction from the WHO Mortality Database

Evgeny M. Andreev

Abstract

The WHO Mortality Database (WHOMD) contains data on populations and deaths classified by sex, age, and detailed causes of death for all countries reporting mortality statistics to the WHO. For many industrialized countries, these data cover a long period starting in the early 1950s. The database includes four huge datasets corresponding to four revisions of the International Classification of Diseases from ICD7 to ICD10. The file format and size make it difficult for inexperienced users to extract data. The WHO itself does not provide a program for extracting data from the database. The Excel spreadsheet WHO-db.xls includes a VBA program permitting the user to extract data from the WHOMD files for a chosen country, calendar period, and single or grouped causes of death. It also allows the user to retrieve data for several countries and years.

Introduction

The WHO Mortality Database (WHOMD) contains detailed cause-of-death data for all countries reporting mortality statistics to the WHO. For many industrialized countries, these data cover a long period starting in the early 1950s. The database includes causes of death classified according to four subsequent revisions of the International Classification of Diseases: ICD7, ICD8, ICD9, and ICD10. Deaths by age, sex, and causes of death are stored in four huge and user-unfriendly text (comma-separated) files. The WHO does not provide a program for extracting data from the database. Only the recommendation that files be imported to Microsoft Access is given as a user support. Access allows for the extraction of data for selected causes, years, and countries, but it does not permit the user to group by causes of death.

The Excel spreadsheet **WHO-db.xls** includes a VBA/Excel program permitting the user to extract data from the WHOMD files for selected country(ies), calendar period(s), and cause(s) of death.

Mortality data by cause of death from the WHOMD can be downloaded from the following WHO page: <http://www.who.int/whosis/mort/download/en/index.html>. After they have been downloaded, the data files should be unzipped.

Table 1 shows the seven files, which are normally downloaded from the WHOMD.

Table 1. The WHOMD files.

File name	Description	File size [#] , Mb
MortIcd7.csv	Deaths numbers by age group, sex, and cause of death classified according to 4 ICD revisions	24.67
MortIcd8.csv		28.55
MortIcd9.csv		79.46
MortIcd10.csv		141.05
POP.csv	Reference populations by age group and sex and live births	1.33
Documentation_*.doc [#]	Brief information on the WHO Mortality Database, detailed file specifications and list of causes of death.	1.61
list_etry_yrs_*.xls	List of available countries and years	0.75

[#] The most recent version for all files is dated 01.12.2009.

All the extracted WHOMD files should be stored in the same folder. The location of the WHO-db.xls file makes no difference.

WHO-db.xls and the Macro

The file WHO-db.xls consists of four worksheets: “Countries,” “CoD,” “Table,” and “Coverage.” Each of these worksheets will be described below.

Worksheet “Countries”

This worksheet serves for selection of countries and years.

Users should specify the full path to the folder that contains all the data in cell B1. The backslash should be the last symbol of the path. For example, users may enter “c:\WHO Data\.” The specified folder will become the working directory for the program.

For selection of counties, required country-codes and years should be specified in columns B and C, respectively.

UN country codes can be found in columns J and K, and also in columns A, C, D, and I of the worksheet “Coverage.” In the latter location, the country code is formed by combining the records in the columns A, C, and D of each row.

Once a country code is entered in a cell of column B, the name of the corresponding country should appear immediately in the neighboring cell of column A. If a non-existing code is entered, the expression #N/A will appear in column A. Users should not use Excel commands like “Cut,” “Insert cells,” or “Delete cells” when entering country-codes in column B. The commands “Copy,” “Paste,” or “Delete” can be used.

Users should insert “All” in cell B4 if they want to retrieve data for all available countries. It is impossible to combine this option with any other country code in column B.

For each country in column A, calendar periods should be specified in corresponding cells in column C. Users can enter single years and/or ranges of years separated by commas. For

example, “1950, 1955-1960, 1971” is a valid expression. All spaces in this list of time periods will be ignored by the program.

If the data for a specified country and/or year are not available, the text “Data are not available,” together with the country name and year will appear.

The country list must begin in cell B3. An empty cell in column B indicates the end of the country list.

The output format should be specified in cell G21. Users who want to save output data for each country-year combination in a separate workbook should enter 1 in this cell. In order to store all output data in one worksheet, 0 should be specified in cell G21.

It is important to note that all output files will be stored in the same working folder.

Worksheet “CoD”

This worksheet is used for selecting of causes of death.

Causes of death must be specified in columns A-F in the worksheet “CoD.” Columns A, B, C, D, and F are used as follows: user-defined cause-of-death numbers (A), user-defined cause-of-death names (B), items of ICD7 (C), items of ICD8 (D), items of ICD9 (E), and items of ICD10 (F). Information about items of the four ICD revisions used in the WHOMD can be found in the WHO-provided file Documentation_*.doc.

For most countries, standard WHO ICD codes are to be used. Some countries, however, have non-standard lists of causes of death. The WHO-provided file list_ctrys_yrs_*.xls allows users to find out which lists of causes of death or which ICD revision correspond to each country-year combination. Items of non-standard lists of causes of death can be then found in the WHO-provided file Documentation_*.doc.

ICD codes in columns C-F should be specified as single terms or ranges separated by commas. For example: “A089, A093-A096” or “E00-E07,E15-E16,E20-E35.” The program ignores spaces in such expressions.

When entering the standard ICD9 and ICD10 codes, aggregate codes must be used whenever possible, rather than series of detailed codes. For example, in order to extract deaths from tuberculosis from ICD9, users should enter code B02 in column E. If this particular ICD item does not exist for some country-year combination, the program will automatically try to find and extract all corresponding four-digit-level items, such as B020 (pulmonary tuberculosis), B021 (other respiratory tuberculosis), B022 (tuberculosis of meninges and central nervous system), B023 (tuberculosis of intestine, peritoneum, and mesenteric glands), B024 (tuberculosis of bones and joints), B025 (tuberculosis of genitourinary system), or B029 (remainder of B02). When working with ICD10 codes, it is sufficient to specify code I in column E for extraction of all deaths from diseases of the circulatory system, or code I2 for extraction of all death from, for example, ischaemic heart diseases.

Users who want to extract data for more than one country should be aware that the selected countries may use different lists of causes of death. In such cases, it is possible to use alternative sets of items from different lists of causes of death. The alternatives must be separated by the symbol | (“or”). The standard ICD codes must be written after the last

separator |. For example, the user might want to use simultaneously the codes of the ICD9 Basic Tabulation List (BTL), as well as the standard items of ICD9 and non-standard codes for one of the newly independent states of the former USSR. In this case, the cause-of-death group “Tuberculosis” can be described as “S020|B02.”

More information about causes of death is available in the WHOMD, and can be found in the following WHO-provided files: Documentation_*.doc, and also in files icd-7_a-list.txt, icd-7_b-list.txt, icd-8_a-list.txt, icd-8_b-list.txt, icd-9_btl-list.txt, icd-10_3char-list.txt, and icd-10_4char-list.txt. These files are provided together with WHO-db.xls.

Running the program

The VBA/Excel program for data extraction from the WHOMD can be executed by clicking on the “Extract” button in the worksheet “Countries.” During the data extraction process, it is not possible to do anything else in Excel. The program’s work would slow down if any other application were to run simultaneously.

Execution of the program can be interrupted by pressing Ctrl-Break.

Program’s output

If the output data are to be placed in one table (value 0 in cell G21 in the worksheet “Countries”), the extraction results will be written in the worksheet “Table.” The data output table includes age-specific population sizes, live birth numbers, and age-specific deaths for all selected causes of death or groups of causes of death.

The program also saves the worksheet “Table” as a separate workbook in a file named OutputWHO-db1.xls in the same working folder. For a majority of data extracts, the data output will not exceed size of a standard Excel worksheet (64,000 rows).

If the output data are too large for one worksheet, the program stores the remaining data in additional output files named OutputWHO-db1.xls, OutputWHO-db2.xls, etc. After each data extraction, the output files should be renamed. Otherwise, these files can be overwritten during the next program’s run. However, Excel will always provide a warning and ask the user whether he or she really wants to overwrite the existing file. In the case of multiple output files, the worksheet “Table” in WHO-db.xls will contain data from the last output.

If the output data are to be placed in separate files (value 1 in cell G21 in the worksheet “Countries”), every country-year combination will be stored in a separate workbook. Corresponding filenames are formed from the UN country-codes and calendar years. For example, filename 2450-55.xls means USA in 1955, and 3160-02 means Japan in 2002.

Worksheet “Table”

As explained above, this worksheet is used for data output.

Death numbers and population sizes for males are displayed in columns A-AH, while those for females appear in columns AI-BP.

Columns A and AI contain the ICD revision; columns B and AJ contain the country codes, and columns C and AK contain the years. The type of age scales (see more in Annex 1 in the WHO-provided file Documentation_*.doc) is placed in columns D and AL. The user-defined names of causes of deaths or groups of causes of death are placed in columns E and AM.

Columns F and AN specify the sex: 1 for males and 2 for females. Columns from G to AD and from AO to BO contain the extracted data.

Columns AH and BP contain all cause-of-death codes corresponding to causes of death that have been extracted from the WHOMD.

If no data are available corresponding to the specified selection, the following message will appear “Data are not available.”

New WHO data and updating of country lists

Several times a year, the WHO provides updates of the WHOMD. The new data files and supplementary files (Table 1) can be downloaded from the WHO website. All other files should be updated as well (list_etry_yrs_*.xls and Documentation_*.doc)

Sometimes (quite rarely), the WHO modifies the list of available countries and regions (the WHO-provided file list_etry_yrs_*.xls. It is therefore advisable to update country lists in the WHO-db.xls from time to time.

The columns “A:H” of the worksheet “Coverage” are copies of the corresponding columns of the file list_etry_yrs_*.xls (under link “availability_mortality” on the WHO website). The updating can be done by clicking the button “Update lists of countries” in the worksheet “Coverage.”

It is possible that the WHO will change the format of this file. In this case, it will be necessary to copy the columns Country; name, Admin1, SubDiv, Year, List, ICD, Updates, starting with the eighth row from this file to the respective columns in the worksheet “Coverage.”

Next, users should start the same macro by clicking the button “Update lists of countries” in the worksheet “Coverage.”

|

Table 2. Columns of the WHO-provided country list

Country	Country code
Name	Country name
Admin1 #	Specified region/Category relevant to each country.
Subdiv #	Category of data (Survey, Reporting Areas, Part, Selected Urban and Rural Areas, Selected Rural Areas, Selected Urban Areas, Cities, Certified Deaths).
Year	Year to which data refer
List	List of ICD revision used (see Documentation...)
ICD	ICD revision
Updates	“New” means that this dataset is included for the first time

If both fields ‘Admin1’ and ‘Subdiv’ are blank, data reported refer to the whole country.

The macro will also fill column I “Complete name” by combing names of region (subpopulations) and/or category of data.

Questions related to working the program in WHO-db.xls should be addressed to Evgueni Andreev

References

World Health Organization. Department of Health Statistics and Informatics. WHO Mortality Data base documentation (1 December 2009 update)