open health data

de Roos, A 2015 A Relational Database of WHO Mortality Data Prepared to Facilitate Global Mortality Research. *Open Health Data,* 3: e1, DOI: http://dx.doi.org/10.5334/ohd.ao

DATA PAPER

A Relational Database of WHO Mortality Data Prepared to Facilitate Global Mortality Research

Albert de Roos[1]

[1] Independent Researcher, Syncyte BioIntelligence, Feike de Boerlaan 255, 1019 WG Amsterdam, The Netherlands
albert@albertderoos.nl

Detailed world mortality data such as collected by the World Health Organization gives a wealth of information about causes of death worldwide over a time span of 60 year. However, the raw mortality data in text format as provided by the WHO is not directly suitable for systematic research and data mining. In this Data Paper, a relational database is presented that is created from the raw WHO mortality data set and includes mortality rates, an ICD-code table and country reference data. This enriched database, as a corpus of global mortality data, can be readily imported in relational databases but can also function as the data source for other types of databases. The use of this database can therefore greatly facilitate global epidemiological research that may provide new clues to genetic or environmental factors in the origins of diseases.

Keywords: mortality; causes of death; World Health Organization

1. Overview

Introduction/Study Description

The WHO Mortality database contain a wealth of information about causes of death for many countries over the last six decades [1, 2, 3]. It gives insight in the age distribution for the various causes of death over the years and includes detailed mortality age distribution. Detailed information about mortality rates for different countries can reveal information about different origin of disease [4, 5]. The combination of the mortality data with other epidemiological and demographical data, for instance on smoking habits or cholesterol levels, can indicate the potential environmental or genetic background of disease [6, 7]. Thus, insight in mortality data can yield important information about national and global mortality trends and may aid in the development of health strategies to target disease.

Even though the benefits of data mining worldwide mortality to discover trends and generate long-term strategies to reduce mortality on major diseases can be great, publicly available data on mortality is not easily accessible for data mining. Research comparing different geographies over time have been sparse and underlying data difficult to reuse. The mortality data as provided by the WHO can not directly be used for data mining and querying, since it needs extensive technical work and analysis. Mortality rates have to be calculated from the raw data in order to compare them. Suitable ICD-code lists that encompass all ICD-versions used over the years have to be made. The availability of an easy-to-use relational database that would relieve researchers from most of the technical hurdles would greatly enhance the exploitation of WHO mortality data for epidemiological research.

In this Data Paper, the WHO mortality data is transformed into a corpus of mortality data in a standard relational database format that allows for easy data mining. The set includes corresponding population data, calculated mortality rates and an ICD-code reference table encompassing all years of ICD registration. The database can be downloaded and imported into a relational database or be combined with other epidemiological or demographic data. The improved ease of access to these data for researchers may be of great benefit for research into global trends and causes of death.

2. Context

Spatial coverage

Global per country. There are 148 different country codes in the final mortality rates tables.

Temporal coverage

Per country over the years 1950–2010 with different time spans per country. The median time span is 26 years, the highest number of years is 62 years (1950–2011) for Japan, The Netherlands and Norway.

Species

Human population

3. Methods

The WHO publishes mortality data in separate text files each containing the different ICD-versions used. Population data is also available corresponding to the same countries and years as in the mortality files. Reference ICD-data is provided in a Word document for all versions of ICD-9 and lower (earlier), while ICD-10 coding can be acquired as a separate dataset. Tools used were MySQL Server (stand-alone) with MySQL Workbench to import, transform and export files. Standard spreadsheet (Excel) and text editing programs (UltraEdit) were used to perform text-based replacement and create transition csv files for import into MySQL. MySQL Workbench was also used to perform the SQL queries and scripts.

Steps

The objective of this study was to generate a relatively simple and transparent dataset suitable for data mining to research mortality rate data based on parameters such as country, years, age and cause of death. **Figure 1A** show the steps that were taken to generate the different mortality, population and mortality rate datasets. From the

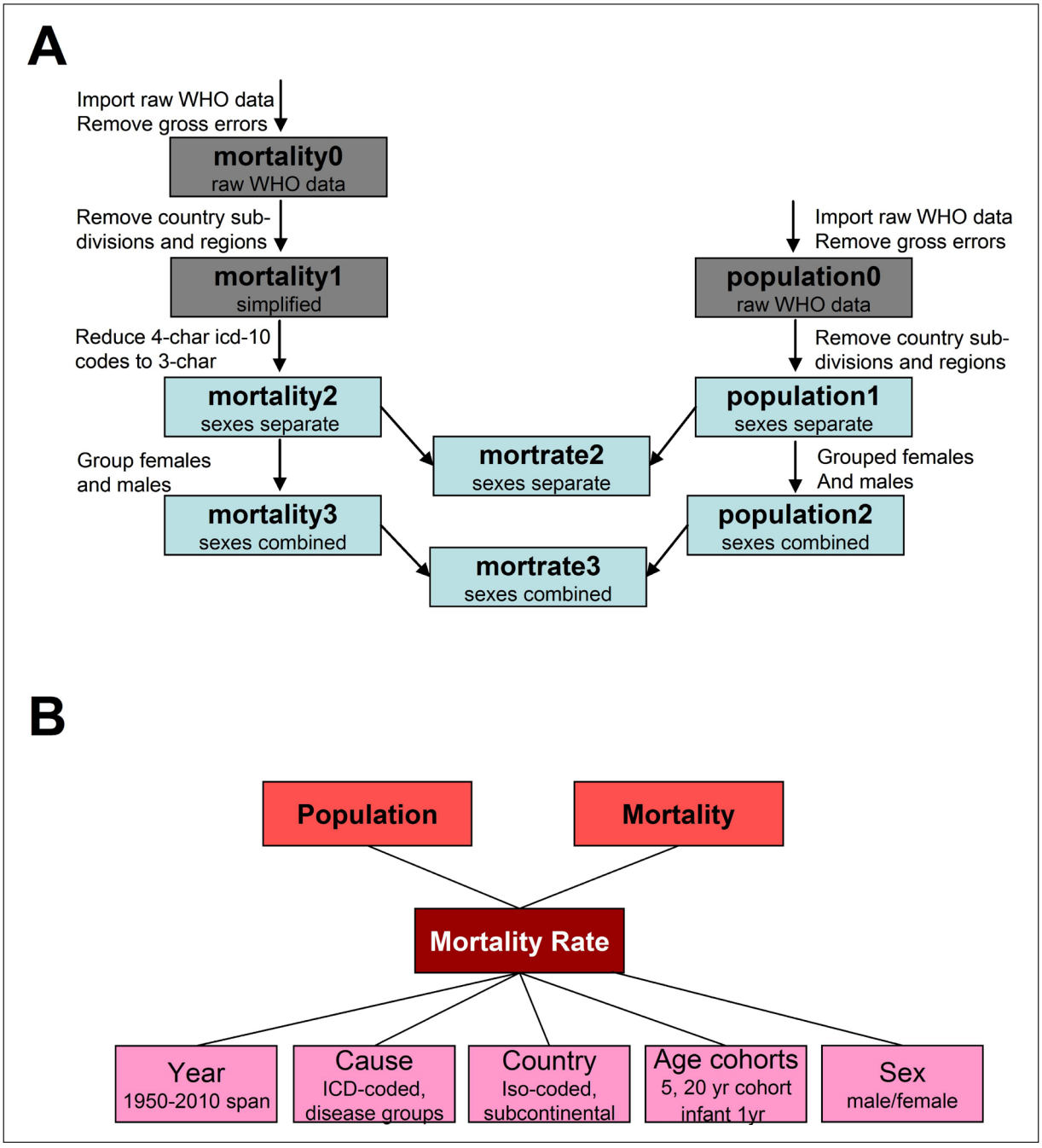

Figure 1: A. Flow schema for importing and preparation of the final databases. In several steps, the raw data in the WHO files were converted to a set of 'production' databases that can be used for mortality studies. The datasets in blue boxes were mainly used in this study. B. Conceptual data model for the mortality rate datasets. Mortality rates were calculated from the population and mortality data. Within the mortality dataset, one can look at the data from different perspectives using the variables year, cause of death, country, age cohort and sex.

original mortality data, new tables were generated where subsequently a) data on regions within countries were removed, b) the causes of death by ICD-10 codes were grouped (see under Sampling Strategy), c) data sets were generated where both sexes were combined, and d) mortality rates were calculated from mortality numbers and population size tables. This setup of the database makes it easy to compare mortality rates between countries but still allows for a transparent data conversion history and ability to see the original (raw) data.

The conceptual data model is shown in **Figure 1B**. The mortality rate database consists of the mortality and population datasets and can be queried over the entities time (calendar Year), the cause of death (ICD-coded), the geographic location (countries and subcontinent), the age cohort (1 year and 5 year) and the sex (female or male). Extra 20-year cohorts were added to facilitate discovering trends.

Sampling strategy

In the process of generating the data sets, several choices were made in order to make the data more easily accessible for data mining. In general, detailed data that is of little use for the comparison of global mortality data over large time spans was not included in the final datasets. In different steps, country subdivisions were removed, ICD-10 codes for detailed causes of death were grouped, mortality rates were calculated and a dataset was generated where both sexes were combined. The following fundamental choices were made when creating the datasets:

a) The creation of a dataset without the subdivisions and regions within a country was done because they were too detailed for most research objectives. Keeping them would also make queries more complex by needing to differentiate between the country and the region.
b) The 4-character ICD-10 coding describes more than 10.000 causes of death while older ICD-coding used over the years contains much less detail. For general research to find global trends in causes of death the 3-character ICD-coding would suffice. Therefore, a separate data table was created where the mortality rate numbers of 4-character ICD-10 codes were grouped into their corresponding 3-character disease group (see **Figure 2**). This led to a reduction of the number of ICD10 codes used from 12,231 4-character to 2049 3-character codes.
c) Mortality can only be compared when the mortality rates are calculated, i.e. by dividing the mortality numbers by their corresponding population size. Therefore, all the mortality rates were pre-calculated in separate tables. This will make data mining queries less complex and more easily available to researchers without advanced SQL skills.
d) The difference between female and male mortality is often too important to ignore, but combined data of both sexes is often sufficient for trend detection. To make the analysis of large datasets easier in those cases, mortality rates were also calculated for both sexes combined.

Quality Control

Tests for the maintenance of data integrity were performed at each step of the transformations. The number of imported and exported rows was verified and basic table data was manually checked for each step. For the mortality rate calculations, representative queries were made that contained the individual mortality and population data and the rate was also manually calculated. Finally, the general rates that were calculated were compared with published data to prevent gross errors in either the raw data or calculations.

A detailed description of the quality control and results can be found accompanying the dataset (Quality Control Mortality Datasets_28122014.xls) in the repository [http://dx.doi.org/10.7910/DVN/28948]. Also all scripts to generate the tables and transform the data are available in the repository (Table Creation WHO_mortality.zip).

Constraints

Not applicable.

Privacy

Not applicable, contains only aggregated information on population level.

Database Table	Country	Year	Sex	ICD-code	Cause-Code	Description	# of deaths
mort1	4210	2000	1	104	C340	Malignant neoplasm: Main bronchus	26
mort1	4210	2000	1	104	C341	Malignant neoplasm: Upper lobe, bronchus or lung	164
mort1	4210	2000	1	104	C342	Malignant neoplasm: Middle lobe, bronchus or lung	8
mort1	4210	2000	1	104	C343	Malignant neoplasm: Lower lobe, bronchus or lung	54
mort1	4210	2000	1	104	C348	Malignant neoplasm: Overlapping lesion of bronchus and lung	9
mort1	4210	2000	1	104	C349	Malignant neoplasm: Bronchus or lung, unspecified	6031
							6292
mort2	4210	2000	1	104	C34	Malignant neoplasm of bronchus and lung	6292

Figure 2: Samples from two different database tables illustrating how the conversion of 4-character ICD-10 codes to a 3-character-code was performed. Data is shown for country The Netherlands in 2010 for male deaths. The detailed entries (C340–C349) for lung cancer in the database table mort1 were combined into one lung cancer group (C43; malignant neoplasm of bronchus and lung) in the database table mort2.

Ethics
Please refer to the terms and conditions of the WHO as listed on their website for the description of the mortality database that was used in this study.

4. Dataset description
Object name
The database that is presented here can be imported using a SQL database dump, named Dump20141228.zip in the repository.

Data type
Secondary data, processed data.

Ontologies
International Classification of Disease, ISO country coding.

Format names and versions
Main format is SQL (database dump and queries). The central idea is that the SQL dump is unzipped and imported in a relational database that can be used directly for querying. The queries used for import, export and transformation are also provided on SQL format. Supporting material is in various data formats including 'MS Office 'office' forms (.doc, .xls, .ppt, .txt).

Creation dates
The mortality database was created in 2014 and used the WHO raw mortality data downloaded from the WHO website on 12 April 2014.

Dataset creators
The author of this article, A.D.G. de Roos, created the datasets described in this article from the original mortality data as created by the World Health Organization (WHO).

Language
English.

Programming language
Main import and export scripts, transformation scrtipts, and queries were performed using SQL in MySQL Workbench.

Licence
The database is free to distribute, adapt and build upon, but restricted by the terms and conditions of the WHO as listed at their website. From their website: Material drawn from the MDB for publication must be accompanied by an acknowledgement of WHO as the source and a disclaimer crediting analyses, interpretations or conclusions to the author of the published data and not to WHO, which is responsible only for the provision of the original information. It should be noted that these data are transmitted on the understanding that no use will be made of them for commercial purposes and that no such permission or right to use may be implied thereby. and is for non-commercial use only. ICD-10 users should register for non-commercial and research use of ICD-10 at the WHO website.

Accessibility criteria
All datasets are limited by the need to accept the policies of use from the WHO. The mortality database contains all the source data files and separate imports that can be directly loaded into MySQL server using standard restore functionality in MySQL Workbench. The data can also be imported into other relational databases or exported in other formats using My SQL server. All SQL import and transformation scripts that were used to generate the databases as well as the SQL queries for the data presented are available.

Repository location
de Roos, Albert, 2015, "WHO Mortality database", http://dx.doi.org/10.7910/DVN/28948 Harvard Dataverse Network [Distributor] V1 [Version].

Publication date
The dataset was published in the repository on 31/01/2015.

5. Reuse potential
For a researcher that uses data mining, it is essential that not only the raw data and its limitations can be understood, but also that the queries can be made transparent and can be tested for accuracy. The dataset that was generated makes it easy to query and drill down on mortality data. The dataset can be easily downloaded and imported directly in a relational database. The database can also be used as a data source for other types of databases or by using natural language query tools but can also be included in a Hadoop cluster.

The simplified data sets and mortality rate calculations offer an attractive start for researchers and the use of this set removes the steep learning curve both in the use of bioinformatics tool as in the understanding of the conceptual data model and its limitations. The mortality presented database can facilitate mortality research and may give new insights in the trends of major diseases worldwide and help to define new strategies for mortality reduction.

Examples of queries illustrating the use of the database are provided in the repository and include historic data on infectious diseases in The Netherlands; breast cancer rates in Japan and The Netherlands; the relation between global smoking habits and lung cancer; and pharyngeal cancers in Hungary versus the Netherlands (Manuscript and Figures.zip).

Competing Interests
The author declares that they have no competing interests.

Acknowledgements
I would like to thank Michel Mutsaers for critically reviewing the manuscript and giving helpful suggestions.

References
1. **Tyczynski, J E, Bray, F, Aareleid, T, Dalmas, M, Kurtinaitis, J, Plesko, I, Pompe-Kirn, V, Stengrevics, A** and **Parkin, D M** 2004 Lung cancer

mortality patterns in selected Central, Eastern and Southern European countries. *Int J Cancer, 109*: 598–610. DOI: DOI: http://dx.doi.org/10.1002/ijc.20019

2. **Viner, R, Coffey, C, Mathers, C, Bloem, P, Costello, A, Santelli, J** and **Patton, G** 2011 50-year mortality trends in children and young people: a study of 50 low, middle and high income countries. *Lancet, 377*: 1162–1174. DOI: http://dx.doi.org/10.1016/S0140-6736(11)60106-2
3. **Ferlay, J, Steliarova-Foucher, E, Lortet-Tieulent, J, Rosso, S, Coebergh, J W W, Comber, H, Forman, D** and **Braya, F** 2013 Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. *Eur J Cancer, 49*: 1374–1403. DOI: http://dx.doi.org/10.1016/j.ejca.2012.12.027
4. **Bray, F, McCarron, P** and **Parkin, D M** 2004 The changing global patterns of female breast cancer incidence and mortality. *Breast Cancer Res, 6*: 229–239. DOI: http://dx.doi.org/10.1186/bcr932
5. **Ott, J J, Ullrich, A, Mascarenhas, M** and **Stevens, G A** 2011 Global cancer incidence and mortality caused by behavior and infection. *J Public Health (Oxf), 33*: 223–233. DOI: http://dx.doi.org/10.1093/pubmed/fdq076
6. **Farzadfar, F F, Finucane M M, Danaei, G F, Pelizzari, P M, Cowan, M J, Paciorek, C J, Singh, G M, Lin, J K, Stevens, G A, Riley, L M** and **Ezzati, M** 2011 National, regional, and global trends in serum total cholesterol since 1980: systematic analysis of health examination surveys and epidemiological studies with 321 country-years and 3.0 million participants. *Lancet, 377*: 578–586. DOI: http://dx.doi.org/10.1016/S0140-6736(10)62038-7
7. **Hosseinpoor, A R, Parker, L A, Tursan d'Espaignet, E** and **Chatterji, S** 2011 Social determinants of smoking in low- and middle-income countries: results from the World Health Survey. *PLoS One, 6*: e20331. DOI: http://dx.doi.org/10.1371/journal.pone.0020331

How to cite this article: de Roos, A 2015 A Relational Database of WHO Mortality Data Prepared to Facilitate Global Mortality Research. *Open Health Data* 3: e1, DOI: http://dx.doi.org/10.5334/ohd.ao

Published: 30 September 2015

]u[*Open Health Data* is a peer-reviewed open access journal published by Ubiquity Press **OPEN ACCESS**