

# Method Protocol for the Human Cause-of-death Data series (HCD@HMD) - Explanatory Notes

## Contents

|  |   |
|--|---|
| Introduction.....                                | 1 |
| 1. Detailed data by cause of death.....          | 1 |
| 1.1 Method of reconstruction.....                | 2 |
| 1.2 Presenting of cause-of-death data.....       | 3 |
| 1.3 Presenting coefficients.....                 | 3 |
| 1.4 Deaths of ill-defined or unknown causes..... | 3 |
| 2. Population exposures and birth counts.....    | 3 |
| 3. Metadata files.....                           | 4 |
| References.....                                  | 4 |

## Introduction

The main inputs for the Human Cause-of-Death Data series (HCD) are country-specific tabulations of death counts by five-year age groups and detailed cause of deaths for national populations. These data, which are initially produced by each country National Statistics Office, are prepared and provided by the HCD country specialists. Usually this is the result of independent research work. Using these death counts and the population exposures from the Human Mortality Database (HMD) we calculate age-specific death rates, crude death rates, and age standardized death rates, including standardized death rates by aggregated age group (0-14, 15-39, 40-64, 65+ years) and by cause of death for four lists of causes of death.

All the output data are provided together with detailed documentation. For each country, there are three blocks of files provided to HCD@HMD users:

1. Detailed data by cause of death
2. Population exposures and birth counts
3. Metadata files.

Whenever authorized by a country National Statistics Office, the raw data (death counts by sex, age group and the cause-of-death categories combined into the HCD shortlist and intermediate list) are published as well on the country page.

### 1. Detailed data by cause of death

The countries for which data series are available in the HCD@HMD all rely on either the World Health Organization International Classification of Diseases or, for historical periods, a Soviet-specific classification. In both cases, the classifications have been periodically revised to account for medical progress, social change, public priorities, and improvements in diagnostics. Cause-of-death data series are severely disrupted by these periodic changes in disease classifications. The existing international databases do not fix this problem. The information provided by the WHO database provides death counts classified according to the International Classification of Diseases (ICD) at the time when the original data were collected. The periodic changes in the ICD thus create discontinuities in cause-specific data series. Eurostat data series are more consistent over time as the data are only provided

for a short list of broad cause-of-death categories, but these data are only available for relatively short periods of time (from 1994 at best) and they do not allow for the analysis of detailed causes of death. Human Cause-of-Death Data series in contrast to other existing databases on causes of deaths, provides harmonized/reconstructed time series with causes of death classified according to a constant (fixed) classification of causes of death taking into account changes in the International Classification of Diseases.

## 1.1 Method of reconstruction

To reconstruct cause-specific death series which are consistent over time, it is necessary to establish transition coefficients between items of two successive ICDs, in order to redistribute deaths classified according to the old classification into items of the new classification. When bridge coding (the double classification of a sample of deaths simultaneously into the old and new classifications) has been performed at a detailed level by the national office in charge of processing cause-of-death data, transition coefficients can be inferred directly from the results, but there are only two countries in the database where this has been done in large enough samples and where the resulting bridge-coding data are available (and only for the last transition, i.e. from ICD-9 to ICD-10), namely England and Wales and the U.S.A. For the other transitions in the two countries and for all transitions in other countries, coherent time series are reconstructed by producing ex-post double coding. The method developed at INED in the 1980s is used as a guideline, but the work was tailored to each country independently. The method is described in detail in the article published by [Pecholdová, Camarda, Meslé and Vallin in 2017](#).

For each classification change, the method comprises three steps (Vallin and Meslé, 1988, 1998; Meslé and Vallin, 1996).

- Setting up one correspondence table which lists, for each item of one classification, all the items of the other classification that are *a priori* equivalent in terms of medical content.
- Building fundamental associations of items that identify the smallest possible number of items containing the same medical contents in both classifications and testing the consistency of the associations over time using statistical testing (Barbieri, Chung, and Boe, 2008; Camarda, Pecholdová, and Meslé, 2015).
- Setting up ex-post double-coding according to the structure of fundamental associations, to calculate the final transition coefficients.

The results derived from the medical logic of the classification rules have to be checked statistically, to detect and solve any remaining breaks in the series. Such checks are carried out visually by age group and sex and by using the statistical tools developed within the HCD team for this purpose (Barbieri, Chung, and Boe, 2008; Camarda, Pecholdová, and Meslé, 2015).

In addition, National Statistical Offices occasionally implement changes independent of the official revisions of the classification by the World Health Organization, in charge of maintaining the schemes. To address this problem, statistical continuity of the series over time is systematically verified and any artificial disruption dealt with appropriately.

Finally, country- and time-specific methods are used to deal with ill-defined causes (Ledermann, 1955; Vallin and Meslé, 1988).

## 1.2 Presenting of cause-of-death data

Causes of death are classified according to four lists: short, intermediate, long, and full. The full list is produced using the 2016 revision of ICD-10 as a standard (though for special cases it is country-specific), while the short, intermediate and long lists are the same for every country. For each list, we provide the same set of indicators (death counts, age-specific death rates, and, for the reconstructed series only, crude death rates and age-standardized death rates). However, in the countries for which we have not received permission to reproduce the original death counts, those are not published for the long and full list.

All data are available in comma-separated (.csv) files. The first columns of the standard files include information on the country, year, sex, the cause-of-death list to which the data apply (short, intermediate, long, country specific full), the ICD code for the underlying cause of death, and (if applicable) the format of the age grouping. For aggregate indicators, the next column provides the corresponding values.

## 1.3 Presenting coefficients

Age-standardized death rates are provided separately using two standards (with one data series for each standard): [the 1976 European Standard Population and the 2013 European Standard Population](#). For age-specific data, the other columns display 26 numbers presenting data by 5-year age groups. Data are provided for all ages up to four open age intervals: 85+, 90+, 95+, and 100+. If the original data are available only up to age interval 85+, the columns corresponding to all older age groups are filled with missing values (dots). If data are available up to age 100+, values are provided for each five-year age group up to 95-99 years and for all open age intervals 85+, 90+, 95+ and 100+ to facilitate comparisons across countries.

Detailed information about the formats of the respective data files can be found in the [Data Formats](#) documentation.

## 1.4 Deaths of ill-defined or unknown causes

A major difference between the reconstructed and unreconstructed data series is that death of ill-defined or unknown causes are redistributed in the reconstructed series but they are provided in separate (ill-defined) categories in the unreconstructed series. The ill-defined categories are thus excluded from the death rate, crude death rate and age-standardized death rates in the reconstructed series. Redistribution of deaths from ill-defined and unknown causes is carried out using the country-specific methods described in each country Background and Documentation file.

Because we realize that some users might be interested in this information for the reconstructed series, a stand-alone file available on each country page provides the numbers of deaths by age group and sex for the three ill-defined categories in the long list, two categories in the intermediate short list and one ill-defined category in the short list as well as for 88 categories in the full list where applicable.

## 2. Population exposures and birth counts

In most countries, the population exposure and birth data used to estimate the denominators of the mortality rates are taken directly from the corresponding all-cause series in the Human Mortality Database (HMD). All rates are automatically recalculated with new population and birth data whenever these are updated in the HMD. However, when HMD population exposures are updated, HCD updates may be delayed for a while. Please note that for some country/years, there may be discrepancies with HMD data as infant deaths and the corresponding birth counts are adjusted for under-reporting. This information is provided in the country Background and Documentation file. The structure of the population and birth data files is similar to that of the cause-of-death data files.

Detailed information about the format of the respective data files can be found in the [Data Formats](#) documentation.

## Metadata files

There are seven documentation files (provided as PDF files) describing the data:

1. *Background and Documentation*. This file provides important information about data preparation, the sources of the original (raw) data, the statistical system of the country, and all other information relevant for understanding the HCD data series.
2. *Reference file*. This contains an overview of the original data sources.
3. *Standard full list of causes of death* (1643 causes of death at the 3-digit level, including 88 ill-defined causes). If a country classification or coding system makes it impossible to apply the standard full list, a country-specific list is applied and provided as part of the documentation.
4. *Standard long list of causes of death* (209 groups of causes, including 3 ill-defined categories in the original data series; 206 groups of causes in the reconstructed series).
5. *Standard intermediate list of causes of death* (58 groups of causes, including 2 ill-defined categories in the original data series; 56 groups of causes in the reconstructed series).
6. *Standard short list of causes of death* (17 groups of causes, including 1 ill-defined category in the original data series; 16 groups of causes in the reconstructed series).
7. *Standard list of ill-defined categories* for all four lists of causes of death.

## References

- Barbieri, M., Chung, R., & Boe, C. (2008). Automating the redistribution of deaths by cause over ICD changes. Second Human Mortality Database Symposium, Max Planck Institute for Demographic Research, Rostock, Germany, 13-14 June 2008.
- Camarda, C.G., Pechholdová, M. & Meslé, F. (2015). Cause-specific senescence: classifying causes of death according to the rate of aging. 80th Annual Meeting of the Population Association of America. San Diego (USA), May 2015. <http://paa2015.princeton.edu/uploads/153074>
- Ledermann, S. (1955). La répartition des décès de cause indéterminée. *Revue de l'Institut international de statistique*, 23 (1–3), 47–55.
- Meslé, F., & Vallin, J. (1996). Reconstructing long-term series of causes of death. *Historical Methods*, 29 (2), 72–87.
- Pechholdová, M., Camarda, CG., Meslé, F. et al. Reconstructing Long-Term Coherent Cause-of-Death Series, a Necessary Step for Analyzing Trends. *Eur J Population* 33, 629–650 (2017). <https://doi.org/10.1007/s10680-017-9453-1>
- Vallin, J., & Meslé, F. (1988). Les causes de décès en France de 1925 à 1978 (Travaux et Documents, No.115, 608 p.). Paris: INED/PUF.
- Vallin, J., & Meslé, F. (1998). Comment suivre l'évolution de la mortalité par cause malgré les discontinuités de la statistique. Le cas de la France de 1925 à 1993. In G. Pavillon (Eds.), *Enjeux des classifications internationales en santé (Questions en santé publique*, pp. 113–156, 220 p.). Paris: Éditions INSERM.