Medicines & Healthcare products Regulatory Agency

CPRD

ONS death registration data and CPRD primary care data Documentation (set 18)

Version 2.1

Date: 09 January 2020

NIHR | National Institute for Health Research

Documentation Control Sheet

Over time, it may be necessary to issue amendments or clarifications to parts of this document. This form must be updated whenever changes are made.

Version	Affected Areas Summary of Change	Prepared By	Reviewed By
1.0	Initial	Shivani Padmanabhan	Rachael Williams, Helen Strongman
1.1	Modified	Susan Eaton	Rachael Williams, Shivani Padmanabhan
1.2	Updated for set 10 Modified data dictionary	Rachael Williams, Dan Dedman	Helen Strongman
1.3	Formatted: new branding	Grant Lee	Sophia Amjad
1.4	Updated for set 12	Arlene Gallagher	Jennifer Campbell
1.5	Updated for set 13	Arlene Gallagher	Shivani Padmanabhan
1.6	Updated for set 14	Rebecca Ghosh	Shivani Padmanabhan
1.7	Updated for set 15 Formatted: new branding	Rebecca Ghosh	Arlene Gallagher
1.8	Updated for set 16 Modified to include CPRD Aurum	Arlene Gallagher, Rebecca Ghosh	Rachael Williams
1.9	Modified data dictionary	Dan Dedman	Rebecca Ghosh
2.0	Updated for set 17	Arlene Gallagher	Rachael Williams, Catherine Bromley (Office for Statistics Regulation), Ben Windsor-Shellard (ONS Head of Lifestyle and Risk Factors Analysis), Elaine Tower (ONS Coding Improvements Manager on Health Analysis and Life Events, Public Policy Analysis)
2.1	Updated for set 18	Arlene Gallagher	Daniel Dedman, Susan Hodgson

Summary of Changes

Version 1.1

- Corrected errors in and added information to the description of the death_matchrank variable
- Incorporated information about lags in registration and potential implications for research use
- Incorporated updated details on the ICD-10 version used by ONS
- Corrected errors in the descriptions of variables cause_neonatal1 through cause_neonatal8

Version 1.2

- Updated for set 10
    - Added information on match rank variable
    - Removed outdated information on multiple matches
    - Updated details of linkage coverage period
    - Added match_rank and dod_partial variables to data dictionary table

Version 1.3

- Updated header and footer to new agency branding

Version 1.4

- Updated for set 12 and with further information on:
  - The linkage coverage period
  - The proportion of patients linked by match_rank
  - The change from ICD-9 to ICD-10 as of 2001 and selection of underlying cause of death
  - The change in causal sequencing from January 2011

Version 1.5

- Updated for set 13 with information on the new coverage period

Version 1.6

- Updated for set 14 with information on the new coverage period

Version 1.7

- Updated for set 15 with information on coverage period and addition of date of registration (dor), gen_death_id, and n_patid_death variables
- Updated header and footer with new agency branding

Version 1.8

- Updated for set 16 with information on the new coverage period
- Updated to include CPRD Aurum
- Updated to include the place of death category indicators

Version 1.9

- Updated data dictionary for [pod_category]: change type to CHAR 255
- Updated data dictionary for [nhs_indicator]: add lookup values

Version 2.0

- Updated for set 17 with information on the new coverage period
- Updated information on the timeliness of death registrations
- Added a section on coding discrepancies

Version 2.1

- Updated for set 18 with information on the new coverage period
- Updated footer with new NIHR branding

ONS death registration data linked to CPRD primary care data

This document provides an overview of the Office for National Statistics (ONS) death registration data, and the available subset that is linked to CPRD GOLD and CPRD Aurum.

What are death registration data?

The Births and Deaths Registration Act (1836) made it a legal requirement for all deaths to be registered from 1 July 1837. The legal requirement to certify and register all deaths occurring in England and Wales means that death registration data provide the most complete information source for mortality statistics. Official mortality statistics for England and Wales are based on the details collected from death registration data.

The registration of deaths occurring in England and Wales is carried out by the Local Registration Service in partnership with the General Register Office (GRO). Information collected at death registration is recorded on the Registration Online (RON) system by registrars. Most of the information is normally supplied by the informant (usually a close relative of the deceased) while the cause of death is usually obtained from the Medical Certificate of Cause of Death (MCCD) completed by a medical practitioner when the death is certified.

When data are entered into RON, there are validation checks to help ensure the details entered are correct. The registrar will also ask the informant to check that the information entered is correct, before the registration is submitted. Regular receipt and diagnostic tests are performed by ONS resulting in weekly contact with the identified registrars to resolve any issues. Once on the ONS database, data are passed through a series of automatic validation processes which highlight any inconsistencies.

Timeliness of death registrations

Deaths should be registered within five days of the date of death. Between 2011 and 2016 there has been a decrease in the timeliness of death registration; in 2016, 61.2% of deaths were registered within five days of the death, compared with 77.7% in 2011 [1]. Deaths considered unexpected, accidental or suspicious will be referred to a coroner who may order a post mortem or carry out a full inquest to ascertain the reasons for the death. The coroner can only register the death once any investigation is concluded and they are satisfied that the death has been thoroughly investigated with a correctly certified cause of death. The time taken to investigate the circumstances of the death can often result in a death registration exceeding the five-day period. While registration delays are commonly only a few days, they can occasionally extend into years. Most deaths are registered within one month (92.4%). Those deaths which have delays in recording are not random but differential by age at death and/or cause of death. The median delay is longer than 5 days for deaths caused by: sudden infant death syndrome (144 days), ill-defined and unspecified causes (141 days), external causes (131 days) and pregnancy, childbirth and the puerperium (128 days) [1].

Cause of death coding

Coding for cause of death is carried out according to the World Health Organization (WHO) International Classification of Diseases (ICD-10) and internationally agreed rules, allowing for international comparisons. Where possible, cause of death – including the selection of underlying and secondary causes - is automatically coded using specialist software, with the remaining deaths being manually coded by highly trained coders. ICD-10 was introduced in England and Wales in January 2001. Since then various amendments have been authorised by WHO. Amendments may (for example) correct errors in the software supporting automatic coding, accommodate new codes in response to new conditions, such as the H1N1 virus (swine flu), or incorporate advances in medical knowledge of the relationship between conditions.

Until December 2010, ONS used the Mortality Medical Data System (MMDS) ICD-10 version 2001.2 software provided by the United States National Center for Health Statistics (NCHS) to code cause of

death. In January 2011, this was updated to version 2010, which incorporated most of the WHO amendments authorised up to 2009.

On 1 January 2014, ONS changed the software used to code cause of death to a package called IRIS (version 2013). IRIS software version 2013 incorporates all official updates to ICD-10 approved by WHO, which were timetabled for implementation before 2014.The ONS provide further details on cause of death coding in section 9 on the ONS website [2]

The accuracy of the automated coding is checked regularly. Cause coding of deaths certified after inquest is performed manually. Completeness checks are conducted to ensure all death registrations have been received. Further checks are also carried out before the annual mortality dataset is finalised.

Impact of coding changes

ICD-10 was introduced in January 2001, replacing ICD-9, which had been in use since 1979 [3]. The Office for National Statistics has carried out a comprehensive study to analyse the changes in mortality statistics that are a result of the change in classification. In ICD-10, the first character of each code is alphabetic rather than numeric. This has enabled the expansion of the number of codes to provide for recently recognised conditions and more detail about common diseases. Some diseases and groups of conditions have been moved between broad groups (ICD chapters), from one to another, to reflect current ideas of aetiology and pathology. These changes mean that data cannot easily be compared across ICD-9 and ICD-10. Some changes in the numbers of deaths attributed to diseases are due to artefacts in the coding system.

In addition to the changes in the coding used there have been several changes to the rules governing selection of the underlying cause of death, reducing the number from 9 to 5. The changes in the application of Rule 3 have had the biggest impact. This rule allows a condition that is reported in either Part I or II of the death certificate to take precedence over the condition selected using the other coding rules if it is obviously a direct consequence of that condition. In ICD-10 the list of conditions affected by Rule 3 is more clearly defined than in ICD-9 and is also broader in scope [4]. The impact of this is to reduce the number of deaths assigned to conditions such as pneumonia and to increase the number of deaths assigned to chronic debilitating diseases. In England and Wales, about 20% of deaths mention pneumonia, so the effect of this rule change is large. Examples of determining sequences and the application of the General Principle and Rules 1, 2 and 3 are available from the WHO [5].

When ICD-10 version 2010 was introduced in January 2011, the ONS conducted a bridge coding study [6]. According to the ONS, the main changes in ICD-10 v2010 were amendments to the modification tables and selection rules used to ascertain a causal sequence and consistently assign underlying cause of death from the conditions recorded on the death certificate.

When IRIS was implemented in January 2014, the ONS conducted an impact assessment. Although 95 % of deaths remained in the same chapter, there were significant increases in the deaths allocated to an underlying cause in some ICD-10 chapters (e.g. the mental and behavioural disorders chapter, which includes dementia) and significant decreases in others (e.g. respiratory disease) [7].

Coding discrepancies

There are a few codes present in the ONS death registration data which may not be found in the ICD-10 Classification. Some codes may be erroneous or are no longer in use (I50.2, J84.2) and others refer to the place of accident in the fourth digit (T58.2 [School, other institution and public administrative area], T71.1 [Residential Institution]). The code U50.9 is used specifically by the ONS to identify deaths involving adjourned inquests. On receipt of the outcome of the inquest, the ONS add a final ICD code relating to the definitive underlying cause. R97 is a code created to identify causes which are synonymous with ‘Cause Unknown’ and is processed differently to other R99 codes (Other ill-defined and unspecified cause of mortality). This code should not appear as the underlying cause of death.

Accessing death registration data linked to CPRD GOLD and CPRD Aurum

ONS death registration data can only be accessed as part of a data extract linked to CPRD primary care data (CPRD GOLD or CPRD Aurum). Access is provided by CPRD subject to MHRA Independent Scientific Advisory Committee (ISAC) approval.

Not all patients in CPRD GOLD or CPRD Aurum are eligible to be linked to death data, for example, due to the region in which they usually resided (outside England), or the lack of a valid NHS number. Source files (linkage_eligibility.txt) are provided to allow researchers to select the subset of patients who are eligible to have a record in the death registration data.

Linkage coverage period

The death registration data includes all deaths *registered* during the coverage period. The latest release (set 18) covers the period from **2nd January 1998 to 31st May 2019**. The date of registration is additionally included alongside the date of death (from set 15 onwards).

Please note that late registration for some deaths means that the proportion of deaths captured is lower for the last year of the coverage period, and this proportion is likely to differ by age at death and cause of death. This is especially pronounced for the last 1-2 weeks of available death data which shows an under count of the total number of deaths as these data do not capture those where the registration of a death has been delayed (e.g. deaths referred to coroners in England, Wales and Northern Ireland, which cannot be registered until investigations have been concluded, and can results in delays of months or years). For more information please refer to the ONS User guide to mortality statistics [2], the ONS analysis exploring the impact of registration delays on mortality statistics [1] and the associated dataset used for this report [8].

Linkage algorithm and the match rank variable

Linkage between ONS death registration data and CPRD primary care data uses an eight-step deterministic linkage algorithm based on four identifiers, shown in Table 1 below. Postcode in the ONS data is based on the usual residence of the deceased as recorded in the death registration data. The linkage is undertaken by NHS Digital, acting as a trusted-third-party, on behalf of CPRD. No personal identifiers are held by CPRD, or included in the CPRD GOLD, CPRD Aurum, or linked death registration data.

Table 1: NHS Digital 8 step linkage algorithm

Step	Match
1	Exact NHS number, sex, date of birth (DOB), postcode
2	Exact NHS number, sex, DOB
3	Exact NHS number, sex, postcode, partial DOB
4	Exact NHS number, sex, partial DOB
5	Exact NHS number, postcode
6	Exact sex, DOB and postcode (where the NHS number does not contradict the match, the DOB is not 1st of January and the postcode is not on the communal establishment list)
7	Exact sex, DOB and postcode (where the NHS number does not contradict the match and the DOB is not 1st of January)
8	Exact NHS number

The matching steps are applied sequentially. If a CPRD GOLD or CPRD Aurum patient record is matched in one step, it is no longer available for matching in subsequent steps. Matching results are summarised in Table 2A and 2B below.

Table 2A: Number and proportion of **CPRD GOLD** patients matched to a patient in death registration data at each step of the linkage algorithm in set 18.

Linkage step (match_rank)	Frequency	Percent
1	618,409	59.3
2	379,275	36.4
3	14,010	1.3
4	12,731	1.2
5	2,043	0.2
6	12,995	1.3
7	2,326	0.2
8	1,722	0.2

Table 2B: Number and proportion of **CPRD Aurum** patients matched to a patient in death registration data at each step of the linkage algorithm in set 18.

Linkage step (match_rank)	Frequency	Percent
1	1,371,610	58.5
2	861,629	36.8
3	32,253	1.4
4	29,871	1.3
5	5,155	0.2
6	32,812	1.4
7	5,639	0.2
8	4,389	0.2

CPRD provides users with a match_rank variable which corresponds to the step at which the match was established. In general, a lower value for the match_rank is considered stronger evidence for a positive match. Note that only patients with a match_rank of 5 or less are considered definitive matches and are included in the linked death registration data. Patients matched on steps 6-8 have been retained in separate files. We envisage that the retained records will primarily be of interest to methodological researchers. *If you are interested in these data, please speak to a member of the CPRD Observational Research team prior to submission of your protocol to the ISAC*.

A minority of patients are linked to multiple death records. These patients are removed from the linked death registration data. However, the data have been retained and are available on request. If you are interested in these data, please speak to a member of the CPRD Observational Research team prior to submission of your protocol to the ISAC.

Data structure and formatting

As far as possible, the linked death registration data is supplied “as is” without any modification or cleaning during processing by CPRD. Where CPRD has modified the data, these are detailed below.

Modification of the coded data: All ICD codes have been normalized into a standard format.

ICD-9 codes: the 1st character of an ICD-9 code is either a number, the letter V, or the letter E (External Causes of Injury and Poisoning). ICD-9 codes will appear in the data with:
3 characters (formatted as XXX)
4 characters (formatted as XXX.X or EXXX)
5 characters (formatted as EXXX.X)

ICD-10 codes: the 1st character of an ICD-10 code is always a letter. ICD-10 codes will appear in the data with:
3 characters (formatted as XXX)
4 characters (formatted as XXX.X)

All codes associated with a death dated from January 2001 have been formatted as ICD-10.

Place of Death: From set 16 onwards CPRD has expanded the information included in the linked ONS death registration data to include a variation of the communal establishment code to differentiate between deaths at home and in hospital. The place of death variable (pod_category) has three categories that provide information on whether the place where death occurred was the home, an establishment (and whether this was local authority or NHS) or elsewhere. An additional integer variable with three categories (nhs_indicator) indicates whether the death occurred within an NHS establishment: 0= elsewhere/at home; 1= NHS establishment; 2 = non-NHS establishment.

Known issues

Before requesting a data extract, you should familiarise yourself with the contents of the ONS death registration data by reviewing the data dictionary as outlined below. Some fields, which are of great potential interest, are/were not mandatory.

- **Date of death (DOD)**: There are some DOD before the start of data collection (1995-1997) and for a small number of records this field is missing, or only a partial date is provided.
- **Date or registration (DOR)**: This field is complete; users may want to consider including information from the DOR when the DOD is missing.
- **Cause of death**: This field is not always complete.

Look-up files

CPRD do not provide ICD-10 or ICD-9 dictionaries.

CPRD recommend acquiring lookup tables for ICD-10 codes from the NHS Digital Technology Reference data Update Distribution (TRUD) Clinical Classifications Service [9] by creating an account, logging in, subscribing to items of interest and downloading the associated files once a subscription is accepted. They can also be contacted via information.standards@nhs.net

ONS death registration data: Data dictionary

1. Patient file (death_patient.txt)

Column name	*Description*	*Type*	*Format*
patid	Encrypted unique key given to a patient in CPRD GOLD or CPRD Aurum [primary key]	INTEGER	20
pracid	Encrypted unique key given to a practice in CPRD GOLD or CPRD Aurum	INTEGER	5
gen_death_id[1]	A generated unique key assigned to a patient in the death registration data. An individual that has contributed data to more than one CPRD practice has the same gen_death_id	INTEGER	20
n_patid_death[1]	Number of individuals in CPRD GOLD or CPRD Aurum assigned the same gen_death_id	INTEGER	3
match_rank[2]	Indicates the quality of matching between a record in death registration data and CPRD primary care data and gives the level of confidence that an ONS death registration record has been correctly matched to a patient in CPRD GOLD or CPRD Aurum.	INTEGER	1
dor	Date of registration of death	DATE	dd/mm/yyyy
dod	Date of death	DATE	dd/mm/yyyy
dod_partial	Partial date of death: where exact date of death is not known, a missing month or day is represented as "00". This field is only used for some patients where the date of death is missing.	CHAR	YYYY-MM-DD
nhs_indicator	NHS establishment indicator for place of death: 0= elsewhere/at home; 1= NHS establishment; 2 = non-NHS establishment	INTEGER	1
pod_category	Indicates a category for the place of death	CHAR	255
cause	Underlying cause of death	CHAR	6
cause1	Cause of death mention ICD10 (non-neonatal deaths only)	CHAR	6
cause2	Cause of death mention ICD10 (non-neonatal deaths only)	CHAR	6
cause3	Cause of death mention ICD10 (non-neonatal deaths only)	CHAR	6
cause4	Cause of death mention ICD10 (non-neonatal deaths only)	CHAR	6
cause5	Cause of death mention ICD10 (non-neonatal deaths only)	CHAR	6
cause6	Cause of death mention ICD10 (non-neonatal deaths only)	CHAR	6
cause7	Cause of death mention ICD10 (non-neonatal deaths only)	CHAR	6
cause8	Cause of death mention ICD10 (non-neonatal deaths only)	CHAR	6
cause9	Cause of death mention ICD10	CHAR	6

[1] Variable generated by CPRD.
[2] An eight-step process is used to match patients in CPRD primary care data (CPRD GOLD or CPRD Aurum) and ONS death registration data using some or all of the following: NHS number, date of birth, sex and postcode. Only data for patients matched using steps 1-5 has been provided.

	(non-neonatal deaths only)		
cause10	Cause of death mention ICD10 (non-neonatal deaths only)	CHAR	6
cause11	Cause of death mention ICD10 (non-neonatal deaths only)	CHAR	6
cause12	Cause of death mention ICD10 (non-neonatal deaths only)	CHAR	6
cause13	Cause of death mention ICD10 (non-neonatal deaths only)	CHAR	6
cause14	Cause of death mention ICD10 (non-neonatal deaths only)	CHAR	6
cause15	Cause of death mention ICD10 (non-neonatal deaths only)	CHAR	6
cause_neonatal1	Cause of death mention ICD10 for neonatal deaths only	CHAR	6
cause_neonatal2	Cause of death mention ICD10 for neonatal deaths only	CHAR	6
cause_neonatal3	Cause of death mention ICD10 for neonatal deaths only	CHAR	6
cause_neonatal4	Cause of death mention ICD10 for neonatal deaths only	CHAR	6
cause_neonatal5	Cause of death mention ICD10 for neonatal deaths only	CHAR	6
cause_neonatal6	Cause of death mention ICD10 for neonatal deaths only	CHAR	6
cause_neonatal7	Cause of death mention ICD10 for neonatal deaths only	CHAR	6
cause_neonatal8	Cause of death mention ICD10 for neonatal deaths only	CHAR	6

References

[1] Office for National Statistics, “Impact of registration delays on mortality statistics, 2016" [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/impactofregistrationdelaysonmortalitystatistics2016. [Accessed 01-April-2019].

[2] Office for National Statistics, "User guide to mortality statistics" [Online]. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/userguidetomortalitystatisticsjuly2017. [Accessed: 30-April-2019].

[3] Office for National Statistics, “Main changes in ICD-10 by chapter.” [Online]. Available: http://webarchive.nationalarchives.gov.uk/20160105160709/http:/www.ons.gov.uk/ons/guide-method/classifications/international-standard-classifications/icd-10-for-mortality/main-changes-in-icd-10-by-chapter/index.html. [Accessed: 23-May-2017].

[4] Office of National Statistics, “User guide to mortality statistics” [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/userguidetomortalitystatisticsjuly2017. [Accessed: 26-April-2019]..

[5] World Health Organisation, “Rules and guidelines for mortality and morbidity coding,” *Int. Classif. Dis. Relat. Heal. Probl. Tenth Revis. Vol. 2*, pp. 31–92, 2004.

[6] Office for National Statistics (ONS), “Results from the ICD – 10 v2010 bridge coding study,” *Statistical Bulletin*, 2011. [Online]. Available: http://webarchive.nationalarchives.gov.uk/20160105160709/http://ons.gov.uk/ons/rel/subnational-health3/results-of-the-icd-10-v2010-bridge-coding-study--england-and-wales--2009/2009/statistical-bulletin--results-of-the-bridge-coding-study.pdf. [Accessed: 23-May-2017].

[7] Office for National Statistics, "Impact of the Implementation of IRIS Software for ICD-10 Cause of Death Coding on Mortality Statistics, England and Wales" [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/impactoftheimplementationofirissoftwareforicd10causeofdeathcodingonmortalitystatisticsenglandandwales/2014-08-08. [Accessed: 26-April-2019].

[8] Office for National Statistics, "Dataset: Impact of registration delays on mortality statistics" [Online]. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/impactofregistrationdelaysonmortalitystatistics. [Accessed: 30-April-2019].

[9] NHS Digital, “Technology Reference data Update Distribution (TRUD).” [Online]. Available: https://isd.digital.nhs.uk/trud3/user/guest/group/0/home. [Accessed: 23-May-2017].