ICD10h: Historic cause of death coding and classification scheme for individual-level causes of death

Manual

Prepared by:
Alice Reid
Eilidh Garrett
Maria Hiltunen Maltesdotter
Mayra Murkens

On behalf of:
The Great Leap (Cost Action)
Studying Health in Port Cities (SHiP and SHiP+ network)
Digitising Scotland

Version 1: July 2024

ICD10h is a research tool created to facilitate the study of historical cause of death records and should not be used for any official purpose. It is based on the International Classification of Diseases, 10th Revision (ICD-10) version 2016 (Geneva: World Health Organization 2016) but is not a recognised version or extension of ICD-10 and is not authorised by WHO. However we have consulted with WHO: they recognise that ICD10h is a useful academic methodology and have not raised any objections to its creation. Data coded using ICD10h are not directly comparable with data coded in ICD-10, and the underlying or primary cause of death derived using the ICD10h methodology may be different from the underlying cause derived in ICD-10 according to the WHO rules. Please note that ICD-10 version 2016 is not the most recent version of ICD-10; and that WHO now recommend the use of ICD-11; a more advanced and detailed classification.

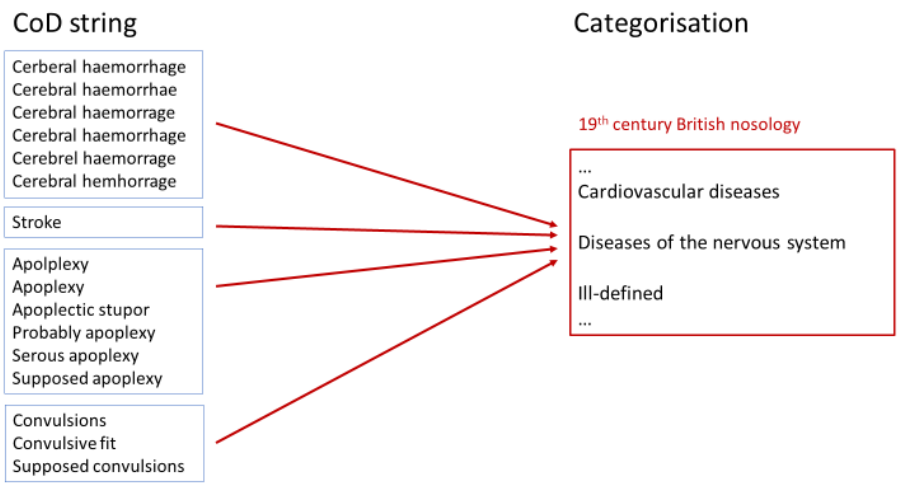
Contents

1. Background and overview of ICD10h[1]

1.1 Previous approaches to individual historic causes of death and justification of our approach

Previously those working with individual level causes of death have allocated each cause of death (COD) string (or cleaned string) straight into one of a small number of COD categories. These categories are often based on a classification or nosology contemporaneous with the period being studied, to enable comparisons with published data for the wider geographic area from which the data in the (usually smaller) dataset was drawn. Some authors adapt contemporary classifications or devise their own to enable particular causes of interest to be singled out, for example the BeRaSaRo-system designed by Bernabeu-Mestre *et al*. which was based on the classifications used by Jacques Bertillon in 1899 and by Thomas McKeown in 1976, the latter itself based on the nineteenth century British nosologies.[2] Figure 1 demonstrates this approach. On the left of the figure are some examples of original cause of death strings, complete with mis-spellings, which can be grouped into four tidier terms: cerebral haemorrhage, stroke, apoplexy and convulsions. In a nineteenth century British nosology, all of these four terms were placed in the 'diseases of the nervous system' category. While this has been a successful strategy for small, time-limited (generally nineteenth-century) data sets, it has several limitations that make it difficult to compare results for different places or long time periods.

Figure 1: Traditional approach to individual-level causes of death: coding straight to a 19th century British nosology



Firstly, such a coding approach prevents flexibility as it necessitates a blanket decision that a specific cause belongs in a particular group. This can be problematic because, as the discussion above has indicated, the use of particular causes changes over time, and may vary from place to place. Secondly, such categorisations are often good for comparisons over short time spans, but make it difficult to compare over long time spans or between places with different death coding contexts. This is particularly the case when

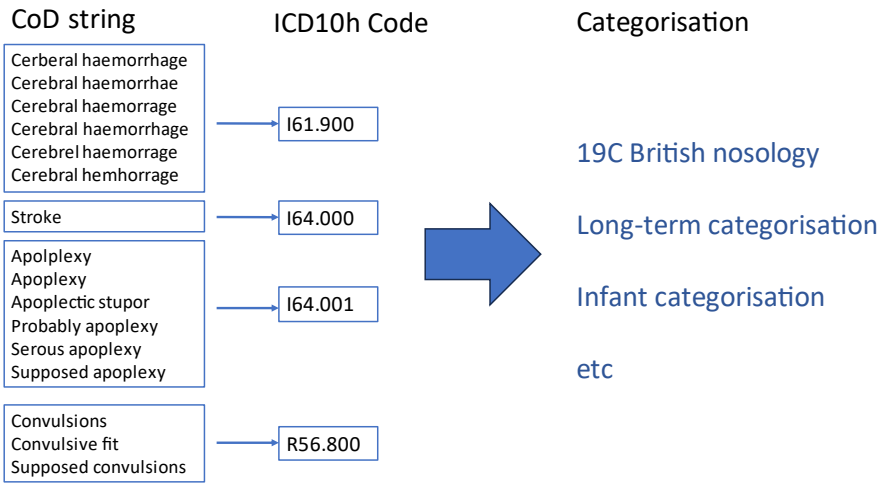
[1] This section is taken from a paper describing the system which is currently being prepared for publication.

[2] J. Bernabeu-Mestre et al., 'El análisis histórico de la mortalidad por causas: problemas y soluciones', *Revista de Demografía Histórica*, 21, 2003, 167–93; B. Revuelta-Eugercios et al., 'Older rationales and other challenges in handling causes of death in historical individual-level databases: the case of Copenhagen, 1880–1881', *Social History of Medicine*, 35, 2022, 1116–39.

it is desirable to compare with causes in published tables. The data for which ICD10h was originally designed stretches from 1855 to 1973 - a period of massive changes in the way that causes of death were conceived, recorded and classified. While a contemporary British classification might work well with the nineteenth century part of this data, and the ICD10 system would be suitable for the final part, neither are appropriate for the entire time span. We need a system which will allow us to both consider long-term changes and to compare short periods with contemporaneously published data. A single classification will not do this. Thirdly, researchers are given no precise instructions about how to allocate individual messy and complicated causes into these higher-level categorisations. For example, 'phthisis' referred to respiratory tuberculosis, but when 'abdominal phthisis' was written on a death certificate should this be treated as 'respiratory tuberculosis', 'abdominal tuberculosis' or something different? This is likely to be a particular issue with vague terms which may represent a variety of different conditions.

Some have tried to deal with this issue by classifying to more than one scheme; for example, Revuelta et al. code both to BeRaSaRo and to ICD10.[3] However coding large volumes of individual strings is time-consuming. The solution that we propose, illustrated in Figure 2, is a flexible and detailed coding system where individual strings are each allocated a code. These codes can then be built up into a variety of different categorisations.

Figure 2: ICD10h approach to individual-level causes of death: separating coding and categorisation



We decided to base our own system on ICD10 because the level of detail in ICD10 means that it is suitable for use at the later end of our period and could relatively easily be adapted for historic causes (the opposite would not have been possible as historic nosologies were much simpler than ICD10).[4] We adapted ICD10 by adding over 3,000 additional codes for archaic causes and providing alternative ways of grouping the causes which are suitable for historic eras.[5] Before we provide details of our own system, therefore, it is helpful to outline the ICD10 system itself.

[3] B. Revuelta-Eugercios et al., Older rationales.

[4] See, for example, the Bertillon classification available at https://curiosity.lib.harvard.edu/contagion/catalog/36-990059378550203941

[5] This approach of creating a historic classification system by starting with a modern one is not unique, as it is also the principle behind the extensively used historical classification of occupations (HISCO) (M.H.D. van Leeuwen et al., *HISCO: Historical International Standard Classification of Occupations* (Leuven University Press, 2002).

ICD10 is a highly granular system, providing over 14,000 distinct codes for diseases or CODs.[6] These causes are grouped into 22 chapters, each referring to a high-level COD category. The English language headings of these chapters are shown in Figure 3, which also gives an impression of the range of codes within each chapter. ICD10 is a hierarchical system, so the letter broadly corresponds to a chapter, and the next two digits refer to a more precise disease group within that chapter. Therefore Chapter 1, Certain Infectious and Parasitic Diseases, includes codes from A00 to B99. Within this chapter, A00 refers to varieties of cholera, A01 to varieties of typhoid and so on. As shown in Figure 3, A00.0 and A00.1 refer to variants of cholera due to different biotypes of the cholera bacterium, and A00.9 refers to cholera which has not been specified as due to a particular biotype. This last code follows the principle laid out in 1864 by James Stark, Scotland's first Superintendent of Statistics, that in a statistical nosology there should always be a 'not further specified' option within each category.[7] Such codes for diseases which are named but 'not further specified' tend to be the most widely used codes in historic contexts.

Chapter	Codes	Chapter name
I	A00-B99	Certain infectious and parasitic diseases
	A00-A09	Intestinal infectious diseases
	A00	Cholera
	A00.0	Cholera due to vibio cholerae 01, biovar cholerae
	A00.1	Cholera due to vibio cholerae 01, biovar eltor
	A00.9	Cholera, unspecified
	...	
	A01	Typhoid and paratypoid fevers
	A02	Other salmonella infections
	...	
	A14-A19	Tuberculosis
	A20-A28	Certain zoonotic bacterial diseases
	...	
II	C00-D48	Neoplasms
III	D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E00-E90	Endocrine, nutritional and metabolic diseases
V	F00-F99	Mental and behavioural disorders
VI	G00-G99	Diseases of the nervous system
VII	H00-H59	Diseases of the eye and adnexa
VIII	H60-H95	Diseases of the ear and mastoid process
IX	I00-I99	Diseases of the circulatory system
X	J00-J99	Diseases of the respiratory system
XI	K00-K93	Diseases of the digestive system
XII	L00-L99	Diseases of the skin and subcutaneous tissue
XIII	M00-M99	Diseases of the musculoskeletal system and connective tissue
XIV	N00-N99	Diseases of the genitourinary system
XV	O00-O99	Pregnancy, childbirth and the puerperium
XVI	P00-P96	Certain conditions originating in the perinatal period
XVII	Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S00-T98	Injury, poisoning and certain other consequences of external causes
XX	V00-Y98	External causes of morbidity and mortality
XXI	Z00-Z99	Factors influencing health status and contact with health services
XXII	U00-U89	Codes for special purposes

Figure 3: ICD10 structure
Source: https://icd.who.int/browse10/2016/en

ICD10 therefore contains elements of both **coding**, in that it assigns a specific code to a particular cause at a granular level, and **categorisation**, in that it groups those codes and causes into much smaller numbers of

[6] WHO | FAQ on ICD, https://web.archive.org/web/20041017011702/http://www.who.int/classifications/help/icdfaq/en/, last accessed 2023/12/16. While ICD10 is an international system, some countries have produced adaptations by adding additional codes. Here we work with the standard version.

[7] Fifth detailed annual report of Registrar General of Births, Deaths and Marriages in Scotland BPP 1864 XVII (3251) xxxv-xxxvii.

similar causes or categories. These categories can be assembled into larger classes for ease of analysis. Most coding systems are part of a classification system, however this does not mean the codes have to be transformed into the in-built classification – they can be grouped in other ways, and we use this principle in our new system, which is outlined in the next section.

1.2 An overview of ICD10h

ICD10h is a two-part coding and categorisation scheme based on ICD10. Where there is an exact match, a string is assigned to an ICD10 code. Additional codes (designated by additional digits) are introduced for historic or archaic terms within existing chapters (e.g. 'consumption' under 'pulmonary tuberculosis'). These codes can then be built up into different categories and classes, for example to match existing nosologies, for specific age-groups, for particular time periods, or for international comparisons. Terms which change meaning over time are given the same code each time that specific term is encountered but that code can then be allocated to the appropriate category for a particular era. Vague terms can be reunited with historically similar causes (e.g. 'pleurisy' with respiratory diseases). This section describes the principles of the coding process, and a later section will describe HistCat, one of the historic categorisations we have developed for use with ICD10h.

Here we set out some important principles of ICD10h:

1. ICD10h takes the 4 alphanumeric digit ICD10 codes as a starting point (e.g. A00.9 for 'cholera, unspecified').
2. We transform all 4-digit ICD10 codes into ICD10h codes by adding two further zeros on the end. Therefore A00.9 in ICD10 becomes A00.900 in ICD10h, and still refers to 'cholera, unspecified'. There are a few ICD10 codes which do not have a digit after the decimal point; examples include A33 Tetanus neonatorum, A34 Obstetrical tetanus, A35 Other tetanus, and B03 Smallpox. For these cases we add three zeros after the decimal point, making A35.000 the ICD10h equivalent of A35 in ICD10.
3. We give each COD string an ICD10h code ending 00 if there is an ICD10 code for that particular string (or a standardised version of that string). For example we code 'cholera' to A00.900 and 'stroke' to I64.000.
4. Importantly we code the *word*, not our interpretation of what it meant in a historic context. In most cases, coding the word and coding our interpretation of it result in the same code; we assume, for instance, that the term 'measles' in the mid-19th century refers to the same disease as 'measles' in the early 21st century. When creating a new code, we placed it within the chapter, code and sub-code which corresponds most closely to the weight of evidence about the condition. Hence, consumption is given a code which groups it with 'respiratory tuberculosis unspecified' (A16.9) even though it might not always have been directly equivalent to that disease. The rationale behind this is that the same word or group of words always has the same code; the classification of that code can be adjusted to reflect any changes in the meanings of the word; individual terms can be reallocated from classifications used at different time periods to produce a time series which takes account of changes in the meaning or usage of certain terms.

   An example of where coding the word and our understanding of it do not lead to the same code is provided by 'teething'. In the first half of the nineteenth century 'teething' allegedly killed a considerable portion of babies in Britain. Out of the 53,157 infant deaths occurring in England and

Wales in 1851, for example, 4400, or 8.3 per cent, were recorded as being due to teething.[8] This cause has been linked to 'weanling diarrhoea'. Infants were often weaned in the second half of their first year when their teeth were also emerging, and both of these might have exposed an infant to a wider range of pathogens – either carried by milk substitutes or feeding bottles, or directly from objects which teething infants are likely to put in their mouth. Neither the weaning nor the teething is likely to have caused the death, but it is possible to see why contemporaries reported them as causes. Although it is generally assumed that the historic cause of death 'teething' is an indicator of diarrhoea, in the spirit of 'coding the word rather than the interpretation' we code it to the ICD10 placement of 'teething syndrome' (K00.7) even though that is under 'diseases of the oral cavity', distinguishing it by adding a historic suffix so it becomes K00.704.[9] We then deal with this in ICD10h by grouping it with diarrhoeal diseases or another group as appropriate for the historical context.[10] This allows the same codes to be allocated to the same cause of death string across time. Keeping terms separate, rather than aggregating them into a code with other terms, allows trends in the separate conditions to be examined to determine whether deaths from one cause always closely mirror deaths from another cause in different places and at different times.

5. Where we come across an historic term which does not match exactly to anything in ICD10 (we refer to such terms as 'archaic'), we create new codes in an appropriate part of the nosology using the additional two digits added to the original ICD10 cause.[11] For example we interpret 'apoplexy' as 'stroke' and therefore we adapt the code for 'stroke' (I64.000) to create a code for 'apoplexy' (I64.001). This allows us to examine the changes from the use of historic/archaic terms to more scientific terms over time, as well as the potential for some categorisations to allocate apoplexy to a different categorisation if a setting is discovered where it does not mean 'stroke'.

6. We also create new codes where we suspect a cause changed meaning over time, to allow the 'new meaning' to be separated from the 'old meaning'. For example, ICD10 contains a code for tetanus: A35. By adding the three additional zeros (because there is no .0 for tetanus), we can create codes for the terms 'lockjaw' (A35.001) and 'trismus' (A35.002). In ICD10, 'trismus neonatorum' is coded under A33, 'tetanus neonatorum', while 'lockjaw' is included under A35, 'other tetanus'. However, historical authorities, such as Payne,[12] equated 'trismus' with lockjaw and trismus neonatorum with 'lockjaw occurring in newly-born children'. We therefore grouped trismus with lockjaw under 'other tetanus' in A35. In most higher level classifications, A33 and A35 will be grouped together, but our codes allow the terms 'lockjaw', 'trismus' and 'tetanus' to be pulled apart, should users wish.

   Similarly, in ICD10, B03 represents smallpox. This disease had been declared eradicated in 1980 but the code was retained in ICD10 for monitoring purposes. In order to better represent eras before

[8] Fourteenth Annual report of the registrar-general (1851) (Registrar General's edition) BPP 1855 98 and 128.

[9] It is very unlikely that 'teething syndrome' caused a death, but ICD10 not only provides codes and classifications for causes of death, but also for diseases and conditions which might not lead to death.

[10] Recent evidence suggests that at least in late nineteenth century Ipswich, it is unlikely that infants whose death was attributed to teething died from diarrhoea (E. Garrett and A. Reid, 'What was Killing Babies in Ipswich Between 1872 and 1909?', *Historical Life Course Studies*, 12, 2022, 173–204.).

[11] This gives us potential for 99 additional codes (01, 02 … 98, 99). Numbers above 10 are rarely used, but are included to provide potential for codes for different languages and to give scope for further expansion.

[12] H. Payne, *A Pocket Vocabulary of Medical Terms (with their Pronounciation), for the use of Registrars, Poor Law Officials, Etc.* (London: Hadden, Best and Co., 1884).

eradication, we introduced codes to represent vaccination status (similar to those in place for other infectious diseases and because some historic nosologies also had codes for these) and to capture variations in the nomenclature associated with the disease. Hence ICD10h, B03.000 represents 'smallpox, (with) no mention of vaccination (on the death certificate)', B03.001 indicates blackpox and B03.002 signifies smallpox where the death certificate specifically states that the deceased had *not* been vaccinated. We have not encountered any certificates which stated that the deceased had died from smallpox despite being vaccinated, but code B03.003 has been created to cover this possibility.

7. Cancers: Cancer is the one causal group which is treated slightly differently in ICD10h compared to ICD10. ICD10 chapter II (Neoplasms) covers letters C and D, with C reserved for neoplasms which have been confirmed as malignant using laboratory tests, and D covering 'benign neoplasms' as well as a small number of codes for neoplasms of 'uncertain or unknown behaviour'. It would have been rare for a nineteenth century cancer diagnosis to have involved a laboratory test, and even if it was this is unlikely to have been mentioned on the death certificate. Therefore, if coding to ICD10, all nineteenth century cancers should be placed in chapter D. We decided to take a different course of action in ICD10h and code our nineteenth century cancers to chapter C. If we had not done this, we would have needed to make a large number of new codes in chapter D in order to contain our historic cancers in similar detail to modern ones (as ICD10 Chapter D contains rather few codes for cancers of uncertain origin and offers a smaller number of parts of the body where cancers can occur). We took the view that any cancer recorded as a contributing to a death was malignant, unless the certificate expressly stated otherwise. However, we did create separate sub-categories under each code for the main terms for cancer which occur in the early civil registers: cancer, scirrhus, sarcoma, epithelioma, tumour, malignant disease, and carcinoma, which receive codes .x01 to .x07 respectively. For completeness, we have provided such terms for all parts of the body, although the type of cancer may now be known not to appear in particular sites.

8. Deaths from external causes (including violence): Just as in ICD10, in ICD10h deaths from external causes can be given two codes, a code from ICD10 chapter XIX (S and T) for the injuries sustained (we refer to these as 'injury codes'), such as a fractured leg or gunshot wounds, and a code from chapter XX (V, W, X and Y) which includes mention of culpability (e.g. accident, assault) or intent (e.g. accidental self-harm, intentional self-harm, intent not known) and of the type of incident leading to the injury (e.g. a fall or a vehicle accident). We refer to these as 'circumstance codes'. Although more than one code can be assigned, as with ICD10, a circumstance code should always be given to each relevant death as this allows intent to be distinguished (where it is clear). Approaches to coding external causes can vary according to how the cause of death strings have been parsed; more detail is given in section 2.4.vi.

1.3 Creating ICD10h

To build the ICD10h codes, we concentrated on creating new codes for archaic causes. To create a dictionary of new archaic causes we used two sources. Our first dataset consisted of COD strings for the 20,115 deaths on the Isle of Skye and the 23,715 deaths in the Scottish town of Kilmarnock, 1861-1901, which had been transcribed as part of a previous project.[13] Our second was 22,024 COD strings relating to

[13] Determining the Demography of Victorian Scotland Through Record Linkage (ESRC RES-000-23-0128) held at the Cambridge Group for the History of Population and Social Structure, University of Cambridge.

93,000 deaths on the island of Tasmania 1838-1899.[14] These sources were chosen because they were most relevant to the Scottish data for which the ICD10h was originally designed. Although the Tasmanian data expand the range of possible causes outside the context of northern Britain, the range is still anglophone and in the nineteenth century many doctors in the antipodes, including those born there and those emigrating, were trained in Britain - usually in Scotland which had the largest medical schools.[15] We did not have access to any cause of death strings for the 20th century period, but we assumed this period would be covered by the combination of nineteenth century terms and the more modern detailed causes captured by ICD10.

The following steps had already been taken by the researchers who produced the original data: all strings were parsed so that separate causes of death were placed in different fields (for first cause, second cause, third cause, etc.), and details of the length of the last illness (included in the Scottish data) separated out. Lists of the first, second, third etc. causes were combined and a further list of unique cause of death strings for single causes were produced. These were then tidied by eliminating spelling variations, removing extraneous spaces, and other standardisation. This standardisation took a slightly different form in the two datasets, so there was still some variation. Unique tidy causes for both datasets were combined to give a total of 19,741 'tidy' strings of single causes of death. These causes of death were hand coded to ICD10h, with the help of the online ICD10 search tool.[16] To understand more esoteric terms we turned to historic nosologies, such as those provided in the Registrar General's reports for England, nineteenth century medical texts,[17] and more recent histories of disease.[18] New terms were created during the coding process and reviewed at the end.

We then automatically allocated codes to each of the causes of death reported for an individual, so that individuals had as many codes as they had causes of death. Because some multiple causes of death have their own ICD10 codes (e.g. B05.2 Measles complicated by pneumonia) we inspected all multiple causes and assigned a single code to reflect this, replacing ICD10h codes where necessary (e.g. for complications of some other common infectious diseases). For multiple causes where more than one code was retained we allocated an underlying cause (the disease or injury which initiated the train of morbid events leading directly to death) following ICD10 rules regarding this, which involve ignoring symptoms and trivial conditions and ordering the other causes by a plausible temporal sequence.[19] We used a version of NCHS's ACME (Automated Classification of Medical Entities) software to assign underlying causes based on the

[14] P. Gunn and R. Kippen, 'Household and Family Formation in Nineteenth-Century Tasmania, Dataset of 195 Thousand Births, 93 Thousand Deaths and 51 Thousand Marriages Registered in Tasmania, 1838-1899', 2008.
[15] A. Crowther and M. Dupree, *Medical Lives in the Age of Surgical Revolution* (Cambridge: Cambridge University Press, 2007), 256.
[16] https://icd.who.int/browse10
[17] J. Copland, *A Dictionary of Practical Medicine* (London: Longman, Brown, Green, Longmans and Roberts, 1833); R. Dunglison, *Medical Lexicon*, 2nd ed. (Philadelphia: Lea and Blanchard, 1839); H. Payne, A pocket vocabulary. We also used Rudy's List of Archaic Medical Terms, available at http://www.antiquusmorbus.com/English/EnglishA.htm.
[18] A. Hardy, '"Death is the Cure of All Diseases": Using the General Register Office Cause of Death Statistics for 1837–1920', *Social History of Medicine*, 7, 1994, 472–92; R. Kippen, '"Incorrect, loose and coarse terms": classifying nineteenth-century English-language causes of death for modern use. An example using Tasmanian data', *Journal of Population Research*, 28, 2011, 267–91; B. Luckin, 'Evaluating the sanitary revolution: typhus and typhoid in London, 1851-1900', *Urban Disease and Mortality in Nineteenth Century England*, (London: Batsford, 1984), 102–19; A. Tanner, The Historic Hospital Admission Records Project (HHARP), *The Historic Hospital Admission Records Project (HHARP)*, 2013. Available at: http://www.hharp.org. [Accessed: 20-Nov-2023]; N. Williams, 'The Reporting and Classification of Causes of Death in Mid-Nineteenth-Century England: The Example of Sheffield', *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 29, 1996, 58–71.
[19] WHO, *International statistical classification of diseases and related health problems 10th revision. Volume 2: Instruction manual.*, Fifth edition (Geneva: WHO, 2016).

ICD10 codes associated with the ICD10h code, however there remained a significant proportion of strings that the software was unable to deal with, and these were resolved manually.[20]

1.4 The HistCat classification for long-term comparisons

The availability of individual causes of death has been hailed as a way of circumventing the problems of nosological changes. To some extent this is true, but there is no getting away from the desirability of comparison with published schema, and it is therefore important to match up with at least some published nosologies. Most analysts will want to compare their data with published aggregate statistics, either to see how their smaller area differs from the 'bigger picture' or to see what they can glean about elements of the registration process (such as insight into the precise decisions about the way that individual COD strings were allocated to particular categories by the official coders in the past). Because the original project for which we created the coding scheme aimed to undertake a comparison over 150 years, we wanted to enable a comparison of the numbers of deaths observed in our data with the distributions published using contemporary nosologies. We therefore generated a classification system based on the lowest level of detail in the Scottish nosologies over the relevant time period. We aggregated these up to maintain useful but distinct categories, combining categories where the most detailed categories in the Scottish nosologies were affected by transfer between categories over time. More detailed descriptions of the process can be found elsewhere.[21] The categorisation was created using published data, and it is generally straightforward to allocate individual deaths into the categories used. However a major issue emerged with deaths described as due to 'debility' or 'weakness'. The published nosologies had categories for 'atrophy and debility' under the general heading of 'perinatal causes', and 'senile debility' under 'old age'. 61 per cent of the deaths attributed simply to 'debility' in the Skye and Kilmarnock data occurred among those aged 60 or over (43 per cent occurred in those aged 70 and over), and 25 per cent in those under age 1. We presume that the Registrars General also used age to allocate cause of death. We also recommend doing this, but we needed a solution for allocating strings to categories in the absence of age. We therefore created a 'debility' category, but we recommend reallocating these deaths based on age: those under age one to perinatal causes, those age 70 and over to old age, and the remainder to ill-defined.

The July 2024 release of ICD10h has a very slightly updated version of HistCat (the only change being the renaming of 'Violence' as 'External causes'). We also provide two versions of a specific classification for use with infant deaths. InfantCat was designed for a comparative analysis of infant mortality in eight European port cities published in a special issue of Historical Life Course Studies.[22] InfantCat2024 incorporates changes inspired by the analyses in the special issue.

---

[20] On ACME see: *https://www.cdc.gov/nchs/nvss/mmds/about_mmds.htm.* Last accessed 13/06/2022, although we used a version downloaded in 2014. Another software for resolving multiple codes is IRIS: *https://www.bfarm.de/EN/Code-systems/Collaboration-and-projects/Iris-Institute/Iris-software/_node.html* Last accessed 13/06/2022.

[21] V. Maiolo and A. Reid, 'Looking for an explanation for the excessive male mortality in England and Wales since the end of the 19th century', *SSM - Population Health*, 11, 2020, 100584; A. Reid et al., '"A confession of ignorance": deaths from old age and deciphering cause-of-death statistics in Scotland, 1855–1949', *The History of the Family*, 20, 2015, 320–44; A. Reid et al., 'A century of deaths, Scotland 1855-1955: a view from the civil registers', *Death in Modern Scotland, 1855-1955: Beliefs, Attitudes and Practices*, (Bern: Peter Lang, 2016), 131–60.

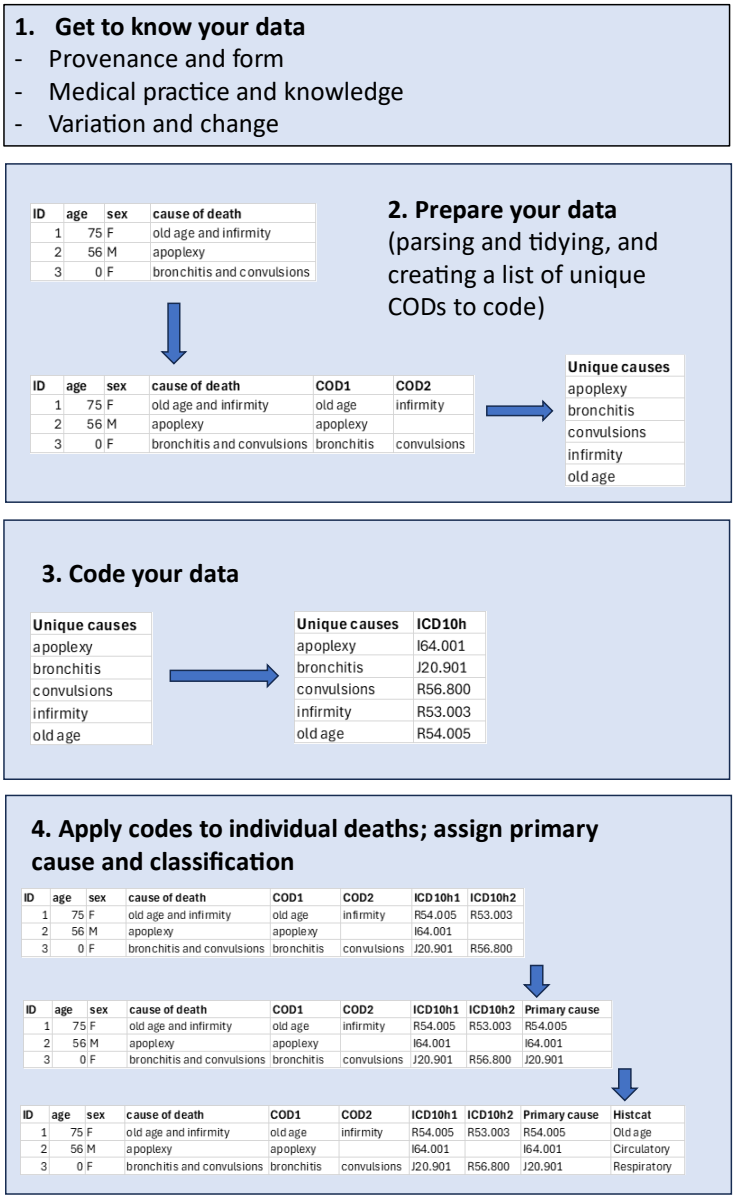
[22] The introduction and overview will be published soon, as well as the two remaining papers, but papers already published are: E. Garrett and A. Reid, What was Killing Babies in Ipswich Between 1872 and 1909?; A. Janssens and T. Riswick, 'What was Killing Babies in Amsterdam? A Study of Infant Mortality Patterns Using Individual-Level Cause of Death Data, 1856–1904', *Historical Life Course Studies*, 13, 2023, 235–64; H.L. Sommerseth, 'What was Killing Babies in Trondheim? An Investigation of Infant Mortality Using Individual Level Cause of Death Data, 1830–1907', *Historical Life Course Studies*, 13, 2023, 61–88; L. Ludvigsen et al., 'Cause-specific infant mortality in Copenhagen 1861–1911 explored using individual level data', *Historical Life Course Studies*, 13, 2023, 9–

2. Instructions for users of ICD10h

2.1 Overview

Coding and categorising deaths can be a lengthy and complex procedure. This section of the guide aims to walk users through the process, following the steps in the following (simplified) diagram:

Figure 4: ICD10h coding and categorisation overview



43; M. Raftakis, 'What was Killing Babies in Hermoupolis, Greece? An Investigation of Infant Mortality Using Individual Level Causes of Death, 1861–1930', *Historical Life Course Studies*, 12, 2022, 205–32; M. Mühlichen and L. Cilek, 'What Was Killing Babies in Rostock? An Investigation of Infant Mortality Using Individual-Level Cause-of-Death Data, 1800–1904', *Historical Life Course Studies*, 14, 2024, 16–40.

2.2 Get to know your data

Cause of death data comes in different forms and formats and this will affect how you parse, code, classify and interpret it. It is highly important to get to know your data, including thinking about the answers to the following questions (sometimes you will not know the answer, but this will still affect how you interpret the data).

Questions about the provenance and form of the data:

- What is the source (burial registers, death certificates or death registers)?
- Who offered the cause of death – a doctor, a priest, the next of kin, or a combination?
- Was the person offering the cause limited to a proscribed list, or could they write anything?
- Has your source been copied from a different source? If so, has all the information been copied and in the same format (e.g. in the Scottish case, we have the civil death registers, with cause of death copied by the registrar from a medical certificate of cause of death (MCCD) supplied by a medical practitioner. The MCCDs had fields for 'primary' (or underlying) and 'secondary' causes but the registers did not).
- Were there instructions to those filling out the source about whether the primary cause was to be listed first or last?
- Who transcribed the data, and what is the potential for further errors to be introduced?

Questions about the medical practice and knowledge:

- What was the level of medical knowledge and the common understandings of disease terms?
- If the cause was always offered by a doctor, were they likely to have seen the deceased during their last illness (if so, they might have a better idea of the cause)? Was an autopsy possible or likely?
- If the cause was sometimes offered by a next of kin or priest, might the deceased have received medical attention?
- Where there any notifiable diseases and what were the potential consequences for the family (and maybe also doctor) if they were reported?

Questions about differences and change:

- Were there differences in any of the factors listed above by region/area/social group?
- How did the factors listed above above change over time?

2.3 Prepare your data

This process reduces the number of different items of data to be coded, plus it reduces the possibility of assigning a different code to the same cause in different instances. How you prepare the data will depend on the form of your original data and on the format of the database into which it has been transcribed. It may also depend on the volume of your data and whether you plan to hand-code it all or to apply a machine-learning approach to coding (in the latter case, standardisation might not be applied). Most of the steps can be achieved using one of a variety of different programmes (e.g. Excel, Access, Python, R) according to the skills and experience of the researcher. Therefore we cannot give you a set of hard and fast rules; instead the following steps are identified as either **necessary** or *desirable*.

2.3.i Transcribe the data (if it is not already transcribed)

- *Transcribe information on COD into a single field, using a unique separator (e.g. /) between individual causes.* Transcribing into a single field will allow you to reassess breaks between individual causes later if necessary. Using a unique separator will make parsing into individual causes easier.
- This step will result in one field, e.g. 'CauseOfDeath' (see Figure 4)

2.3.ii Parse the data

- **Separate different causes of death for the same individual into different fields, while also retaining the original entry in its own field.**
    - *Some information in the cause of death field may not relate to a cause of death, this can also be separated out during the parsing process, however we recommend keeping modifiers such as 'chronic', 'acute', 'perinatal' and so on with the cause of death, as the code assigned may differ for different modifiers.*
    - *Strings where a disease is used as an adjective, such as tubercular meningitis, should not be parsed out into separate diseases (e.g. tuberculosis and meningitis). Tubercular meningitis has a code of its own.*
    - *It can even be helpful to manage the parsing in such a way as to retain useful modifiers in more than one cause of death, particularly in cases of childbirth or accidents and violence. In the case of childbirth, the fact that the childbirth was involved can lead to a more accurate coding of conditions such as haemorrhage, toxaemia, fever and so on. In the case of accidents and violence, it is quite common for both an injury (e.g. liver rupture, severe bleeding) to be given as well as the reason for that (e.g. being run over by a cart). Knowing that such injuries are the result of trauma can lead to more accurate coding.*
    - *Similarly, it is helpful to make sure that in cases where the same disease manifests in different parts of the body (e.g. 'cancer in breast and bones', 'tuberculosis of lung and stomach') that the disease term is kept for individual terms relating each location (e.g. these are best parsed to 'cancer in breast' and 'cancer in bones', and to 'tuberculosis of lung' and 'tuberculosis of stomach').*
    - Parsing can be achieved in a number of different ways:
        - If you have used a unique separator (e.g. /) when entering the data, parsing will be straightforward.
        - You can use a machine guided set of rules to search for separators (e.g. 'and', 'following', 'preceding', ',') but you will need to inspect your data carefully first to identify likely separators, as this will differ by dataset, and afterwards to ensure that you are not splitting causes which should be kept together.
        - You could use a machine-learning approach based on a training dataset.
    - This step will add several new fields, e.g. 'Cause1', 'Cause2', 'Cause3', etc. to data for each individual.

2.3.iii Tidy the data

- *This is generally a useful step which helps to reduce the number of unique strings and lessens the chance that the same string will be allocated different codes. However those with very large data sets who are using machine learning to code their data may not need this step.*
- Create new fields for tidied versions of the parsed data, keeping the original fields intact.

- Identify material which is not a cause of death and separate into a different field unless it is likely to aid in the correct allocation of a cause of death. Examples of material which can be separated out include whether the death was certified by a doctor or the deceased had medical attendance during their last illness and whether the death was referred to a coroner or equivalent. The length of last illness can be helpful in a few cases, such as bronchitis, but generally does not help coding, so codes for the few instances where it is helpful can be adjusted later (see section 2.4.v). Examples of material which is very important to retain with the cause of death because it is so helpful in assigning the correct code include 'acute', 'chronic', and 'congenital'.
- Deaths from violence can take a narrative form and can be very difficult to tidy and parse. We recommend taking extra care not to separate out words which might help allocate the code. Take, for example, the phrase 'drowned in North Sea'. The name of the body of water in which the person drowned is not relevant, but the fact that they drowned in the sea, as opposed to a river or bath-tub is relevant.
- Delete extra spaces between words and at the start and end of strings.
- This step will add a new set of fields, e.g. 'TidyCause1', 'TidyCause2', 'TidyCause3', etc. to the data for each individual.

2.3.iv Create a list of unique CODs

- Combine the fields of tidy causes into a new table, eliminate empty cells and duplicates.

2.3.v Standardise unique CODs

- *This is generally a useful step which helps to reduce the number of unique strings and lessens the chance that the same string will be allocated different codes. However those with very large data sets who are using machine learning to code their data may not need this step.*
- Arrange alphabetically to allow spelling variations to be detected.
- Create a new field for standardised cause of death, keeping the original field intact.
- Use the standardised field to eliminate spelling errors and extraneous information.

2.3.vi Create a new table with just the unique standardised causes of death and eliminate duplicates.

- We do **not** recommend translating non-English terms into English, due to the problems with 'false friends' and ambiguous causes (see below).

At the end of this process you should have a table of unique, tidied, and standardised causes of death, the next stage will allocate a code to each of these. In the process of creating your table of unique causes, you will also have created a series of look-up tables which will enable you to transfer the code for each cause of death back to the original records.

2.4 Allocate a code to each unique COD

2.4.i Overview

Approaching coding in a systematic fashion helps to ensure codes are consistent, (i.e. the specific diagnostic expression always receives the same code) and reproducible (i.e. the same code will be given by different coders, or by the same coder on different days).

Trying to achieve comparable codes between different datasets and across time can be hard enough, but achieving consistency between datasets in different languages comes with additional challenges. One challenge is to decide which causes of death are synonymous with each other, and whether they are

directly synonymous with an English cause, or whether they are unique to a different language. It is often difficult to identify the proper equivalent in English, and coders need to watch out for 'false friends' (words which *sound* the same but do not *mean* the same thing in different languages – see examples of coding non-English causes and a list of false friends in the appendix) which can mean that simply translating a word from another language into English and searching for a cause of death which matches the English word may lead to the wrong code being assigned. Finally, it can be difficult to prioritise between possible alternative codes. Hence the process for assigning codes to individual causes is slightly different for English language and non-English language causes.

A variety of special cases (conditional codes, accidents and violence, cancers, and dagger and asterisk codes) are explained towards the end of this section, but it is easier to understand these after working through the coding steps for more straightforward cases.

The recommended process for assigning codes to unique cause of death strings is slightly different for languages where we have a list of historic strings already coded to ICD10h, which is currently just English. However, when other such lists of coded strings are produced for other languages and published on the greatleap.eu website, the English language process can be followed. Figures 5 and 6 provide workflows to guide users through the process of assigning codes to unique causes of death, for non-English language and English language causes respectively, using the ICD10h tables and helpful resources from the modern ICD10. Sections 2.4.iii and 2.4.iv guide users through these workflows and provide examples. We recommend that you read both of these sections as they each include useful information and approaches irrespective of the language you are working in.

When following the flowcharts, please remember:

- When searching in excel, remember to choose to search by column, and to make sure 'match case' and 'match entire cell contents' are unchecked.
- Do check all possible matches to consider which may be the closest match.
- You are coding the words not your interpretation of the words.
- Avoid codes identified as conditions 'occurring in a disease elsewhere' (see dagger and asterisk codes).
- Usually historic terms fall under 'unspecified' or 'not otherwise specified' (NOS), because few, if any, further details about the cause are given. These causes tend to come at the end of a set of codes referring to a particular disease or condition.
- Look out for qualifiers such as 'infectious' or 'of infectious origin' in the descriptions of the codes as well as in your strings. This is particularly relevant for diarrhoeal diseases (including enteritis, colitis, bowel inflammation and so on). ICD10 places those of non-infectious origin in K52.9, those of infectious origin in A09.0, and those of unspecified origin in A09.9. Most historic causes of this nature will be of unspecified origin and should be coded accordingly.

2.4.ii Resources

ICD10h tables (see also section 5.2 for lists of the fields in each table, and a figure demonstrating the relationships between the tables)

Masterlist

- The definitive list of ICD10h codes with one row for each code, including an historic description of the cause.

- Where the historic code is the same as ICD10 (with the addition of 00) the ICD10 description is used.
- In both the historic and ICD10 descriptions, words in parentheses indicate terms that do not necessarily occur in the cause of death as written, but denote variations which should also be included under that cause.
- The Masterlist also contains the HistCat to which each ICD10h is allocated.

HistoricStringsEnglish

- A list of English language historic causes from the Scottish and Tasmanian deaths used to create the system, together with the ICD10h code and description of the cause.
- For certain external causes, additional columns are provided to allow more detail of injuries to be specified (see section 2.4.vii for more details).
- This table reflects the original strings which have been subject to only minimal tidying and standardisation. It therefore provides a variety of different ways to describe the same cause (e.g. 'stomach.disease' and 'disease.stomach'). This can make it easier to match up with new source material, but it also means there is often more than one row per code.
- Many ICD10h codes do not have an entry in HistoricStringsEnglish – it only contains those encountered in the datasets used to create the system. This list is not exhaustive or definitive – it is purely provided as a helpful guide. Therefore, unless your English language cause is an exact match, you should always check in the Masterlist (and often also modern ICD10 resources) for the cause you are trying to code.
- HistoricStringsEnglish is linked to the Masterlist by the ICD10h code.

Modern ICD10 resources

ICD10 2016 online lookup (https://icd.who.int/browse10/2016/en)

- The online version indicated above has a useful search function.
- The 2016 version of ICD10 has been chosen as the definitive version when working with ICD10h. To maintain consistency and comparability, coders should try **not** to use versions from other years, nor should they use ICD11.
- If your sources are not in English, we recommend trying to find a version of ICD10 2016 in the same language as your records. If you cannot find a 2016 version in your language, a version from another year in your language will be acceptable.

ICD10 instruction manual (link to source)

- Detailed coding rules on which ICD10h rules are also based.
- If in doubt about complex cases, refer to these rules.

ICD10 alphabetical list (link to source)

- Alphabetical index of all the terms included in ICD10.
- Useful for double-checking all the occurrences of a term to identify the correct choice.
- Section I provides an index to diseases and nature of injury; Section II is an index of external causes of injures; Section III contains a detailed table of drugs and chemicals and their placement in ICD-10 depending on the circumstance (poisoning by accident, suicide, undetermined intent) and nature of the injury.

MESH (Medical Subject Headings) dictionaries and translations

The US National Library of Medicine controlled vocabulary thesaurus of medical terms. Background information can be found here: https://www.nlm.nih.gov/mesh/meshhome.html, and can be searched here: https://www.ncbi.nlm.nih.gov/mesh/ or https://meshb.nlm.nih.gov/search. The service also offers dictionaries between English and a variety of other languages:

- Spanish-Portugese-French-English: https://decs.bvsalud.org/en/
- German-English MesH https://www.zbmed.de/en/open-science/terminologies/german-mesh
- French-English MeSH: https://mesh.inserm.fr/FrenchMesh/
- Norwegian-English MeSH: https://mesh.uia.no/
- Swedish-English MeSH: https://mesh.kib.ki.se/
- Greek version of MeSH https://www.nlg.gr/eresources/mesh-medical-subject-headings/

2.4.iii Coding process where no sets of historic strings coded to the most recent version of ICD10h are available (currently all non-English languages) and examples

For non-English language causes of death, we strongly advise against translating your causes into English and coding the English translation. This is because there may be no clear or likely match with an English cause of death. In addition, some causes may be very particular to the local or national context; they may be used differently to their use in other languages (e.g. they might be 'false friends'). It is very important to use a version of the modern (ICD10) in your own language as this can provide the best accepted meaning of the terms. Similarly it is important to identify historic sources in your own language and to familiarize yourself with local medical practices and terminology in use in the relevant time period.

Start by searching the modern ICD10, following the eight basic guidelines from section 3.3 of the ***ICD10 Instruction manual*** (check 1 in Figure 5):

> (1) **Identify the type of statement** to be coded and refer to the appropriate section of the **Alphabetical index**. If it is a disease or injury, consult section I of the index. If it is the external cause of an injury, consult section II.
>
> (2) **Identify the lead term.** This is usually a noun referring to the disease or illness.
>
> (3-4) **Identify all modifiers** (for example acute, chronic, congenital, puerperal, perinatal) and how these affect the code (the alphabetical index is very helpful here).
>
> (5) **Follow any cross-references** ('see' and 'see also') found in the alphabetical index.
>
> (6) Refer to the **tabular list** (ICD10 arranged by code, e.g. in the online lookup) to verify the suitability of the code selected.
>
> (7) Be guided by any **inclusion or exclusion terms** under the selected code.
>
> (8) If a good match is found, a provisional ICD10 code can be identified ($ICD10_1$).

This should then be checked against the historical reference literature (check 2 in Figure 5) to confirm the provisional code ($ICD10_2$) before locating that ICD10 code in the Masterlist and HistoricStringsEnglish (check 3 in Figure 5) to see if any specific historic codes have been assigned ($ICD10h_3$). Sometimes it is the case that these three checks do not yield codes with the same ICD10, and these then have to be resolved in the step 'Prioritise collected information'. For example, the Swedish string 'krupp' can be found in the Swedish ICD10-2016 version under J05.0 (which has the English ICD10 description of 'Acute obstructive laryngitis [croup]'), and thus $ICD10_1$ is J05.0. The historic reference literature, however, suggests that

'krupp' meant Diphtheria, yielding $ICD10_2$ of A36.9. Examination of the Masterlist for J05.0 reveals that a particular historic code, J05.001, has been assigned for 'Croup', whereas although there are also specific codes under A36.9 these just refer to Diphtheria. In this case, the combination of the Swedish ICD10 identifying 'krupp' as J05.0, with English translation 'croup', and the presence of the specific ICD10h code for croup J05.001, indicates that the final code assigned should be J05.001. More examples of resolving preliminary codes for different forms of the Swedish term 'krupp' are provided in Table 1.

Table 1: Examples of resolving preliminary codes for different forms of the Swedish term 'krupp'

Disease	**Match in: Modern ICD-10 2016, Alphabetical index (ICD-10)**	**Medical reference literature, old nomenclatures etc. (Swedish)**	**Masterlist**	**Prioritised ICD10h**
Croup	'Acute obstructive laryngitis [croup]' (also known as pseudocroup) **$ICD10_1$**=J05.0	Croup='Diphtheria' **$ICD10_2$**=A36.9	ICD10h description='Croup' **$ICD10h_3$**=J05.001	J05.001
Croup, diphtheritic	**$ICD10_1$**=A36.2	**$ICD10_2$**=A36.9	ICD10h description= 'Diphtheritic croup' **$ICD10h_3$**=A36.201	A36.201
Croup, false	**$ICD10_1$**=J38.5 (English match) **$ICD10_2$**=J05.0 (Swedish match)	'Falsk krupp' SW = pseudocroup **$ICD10_3$**=J05.0	ICD10h description= 'Laryngeal spasm' **$ICD10h_4$**= J38.500	J38.500

If there is no obvious match in the modern ICD10, a more careful examination of the diagnostic expression is needed to aid searching and identification of possible causes (preliminary ICD10 codes) using the reference literature. As before, a step-by-step process of checking the reference literature, the modern ICD10 and the Masterlist should be carried out, and any discrepancies in preliminary codes resolved in the final step.

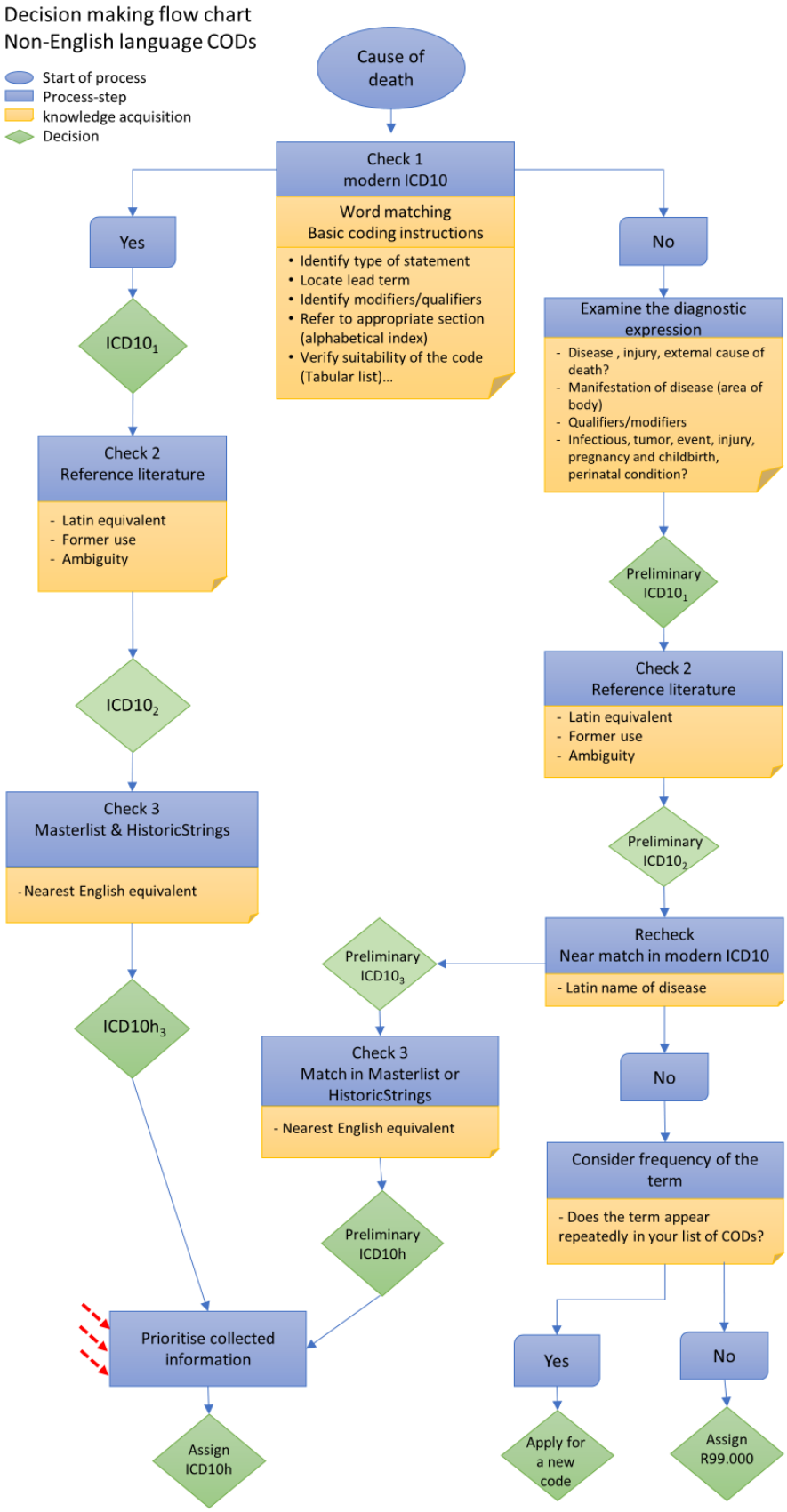
If there is no close match for your string and there are also a sizable number of deaths from that string, you can consider requesting a new code (see section 2.4.ix).

Further examples, from the Dutch language, are explained below and summarised in Table 2.

'*Phthisis*'

Phthisis will not render any results in the modern Dutch ICD10 (contrary to the English ICD10 where it is included in the online search but not the hierarchical list) because it is an archaic term for tuberculosis, predominantly of the lung. The English Masterlist does find an exact match as the Dutch and English both used the exact same term. The code in the Masterlist for phthisis is A16.901, which is the correct choice. If the COD had been given the qualifier pulmonary (we find phthisis pulmonum quite often in the Dutch CODs), the appropriate code according to the Masterlist would instead be A16.903.

Figure 5


Decision making flow chart
Non-English language CODs
Start of process
Process-step
knowledge acquisition
Decision
Cause of death
Check 1
modern ICD10
Word matching
Basic coding instructions
• Identify type of statement
• Locate lead term
• Identify modifiers/qualifiers
• Refer to appropriate section (alphabetical index)
• Verify suitability of the code (Tabular list)...
Yes
No
ICD10$_1$
Check 2
Reference literature
- Latin equivalent
- Former use
- Ambiguity
ICD10$_2$
Check 3
Masterlist & HistoricStrings
- Nearest English equivalent
ICD10h$_3$
Examine the diagnostic expression
- Disease , injury, external cause of death?
- Manifestation of disease (area of body)
- Qualifiers/modifiers
- Infectious, tumor, event, injury, pregnancy and childbirth, perinatal condition?
Preliminary ICD10$_1$
Check 2
Reference literature
- Latin equivalent
- Former use
- Ambiguity
Preliminary ICD10$_2$
Recheck
Near match in modern ICD10
- Latin name of disease
Preliminary ICD10$_3$
Check 3
Match in Masterlist or HistoricStrings
- Nearest English equivalent
Preliminary ICD10h
No
Consider frequency of the term
- Does the term appear repeatedly in your list of CODs?
Yes
No
Apply for a new code
Assign R99.000
Prioritise collected information
Assign ICD10h

'*Longontsteking*' (literal translation inflammation of the lung, the Dutch vernacular term for pneumonia).

In check 1 we check the modern ICD10, in this case the Dutch language version, which gives no result for *longontsteking*. However, the synonym medical term in Dutch is *pneumonie*. Searching for *pneumonie* produces several results. It shows that without further specification about the cause of the pneumonia, it should be allocated the code J18. Checking this with the English Masterlist (check 3), the English equivalent pneumonia (unspecified) is indeed given the code J18.900 in the ICD10h.

'Kinderziekte' (literal translation children's disease)

A first hunch might give the impression that this is an ill-defined description of any disease that could occur in childhood – indeed ICD10h contains a Swedish term which also translates to 'children's disease' but which could refer to a number of conditions. Despite this, the steps in the flow-chart should be followed. The first step is to consult ICD10, but as *kinderziekte* is an archaic term it does not appear. Further examination of the diagnostic expression is needed. We also find the description *kinderziekte gev*, where the *gev* may be an important qualifier. Apparently the *gev* appears more often together with the mentioning of smallpox, making it likely that is means *gevaccineerd* (vaccinated). The next step is to consult contemporary medical dictionaries and handbooks (check 2), and this reveals that the Dutch term *kinderziekte* is an archaic term used to describe smallpox. We then return to ICD10 to locate the general codes for smallpox, B03, and look in the ICD10h Masterlist to see if an appropriate historical code has been created. This reveals that *kinderziekte* has been allocated the code B03.004. This code was created after application for a new code, as the cause of death occurred several times in multiple Dutch datasets.

'*Angina*'

Another example is the case of angina, an ambiguous cause. In the English language, angina is most often used to describe the heart affliction 'angina pectoris' (I02.900). However, the word angina is also part of other, quite different, disease terms, such as Ludwig's angina (K12.2 a form of mouth abscess), Vincent's angina (A69.1, a form of pharyngitis or tonsillitis) and Diphtheritic membranous angina (A36.0, a form of diphtheria). Both the English and Dutch versions of ICD10 state that 'angina – not otherwise specified' refers to 'angina pectoris'. So, in the spirit of coding the word, 'angina', without further specification, should be coded to I02.901. However, it is important to note that consultation of the Dutch medical literature (check 2) reveals that on its own, the term angina was actually an archaic term for inflammation of the throat. Because the codes should reflect the word if there is an exact match in the ICD10, the correct place to account for this archaic use is in the step after coding, the classification (a Dutch version of the HistCat classification, adjusted to place the code I02.901 in a category with other diseases of the throat rather than with heart conditions, is currently under development). Here the collected information, and the consistency in coding across languages, together with the spirit of coding the word, were given priority. The fact that the cause of death has a different meaning in Dutch can be accounted for in a different step, and therefore the other information holds priority over the archaic meaning.

Table 2: Non-English language coding examples

Original string	ICD10h	ICD10h description	ICD10	ICD10 description
phthisis	A16.901	Phthisis	A16.9	Respiratory tuberculosis
longontsteking	J18.900	Pneumonia	J18.9	Pneumonia, unspecified
kinderziekte / gh s/	B03.004	Kinderziekte	B03	Smallpox
Angina	I02.901	Angina	I02.9	Angina pectoris, unspecified

2.4.iv English language coding process and examples

For causes of death in the English language, we already have a coded set of historic strings (HistoricStringsEnglish), and this makes assigning codes to common strings relatively easy. This is because if a particular string has already been assigned a code, that is the code that should be used. Therefore the recommended process starts by checking for exact matches in the HistoricStringsEnglish table (Check 1 in Figure 6) and assigning ICD10h codes for strings where there is an exact match.

If there is no exact match in the HistoricStringsEnglish table, the string (diagnostic expression) should be examined to determine its meaning, key elements, and lead term, to help with further searching. For example the string 'cholera asiatica', which does not appear as an exact match in HistoricStringsEnglish, seems to refer to a particular form of cholera, so the lead term is cholera and 'asiatica' is a term to distinguish a particular form of cholera.[23] However other variations of 'asiatica' could be acceptable, such as 'asiatic' (or possibly 'asian'). The next step is to check for a near match in HistoricStringsEnglish (Check 2 in Figure 6). There are no near matching strings starting with the word cholera, but 'asiatic cholera' is a very near match. This allows a preliminary ICD10h code, A00.900, to be identified.

Many strings will have no near match in HistoricStringsEnglish. Four examples are 'aplastic anaemia', 'green sickness', 'rising of lights' and 'surfeit'. Check 3 involves searching for the relevant string in the modern ICD10 online tool using the eight basic guidelines from section 3.3 of the *ICD10 Instruction manual* as detailed in section 2.4.iii above. In the example of 'aplastic anaemia', the ICD10 code D61.9 represents 'aplastic anaemia, unspecified'. When search for 'green sickness' in the online ICD10 tool, users are directed to the ICD10 code D50.8 ('Other iron deficiency anaemias'). Note that the dictionary behind the ICD10 online tool is clearly much more extensive than the code descriptions, so use of the tool can be very helpful even for historic terms. Neither of the terms 'rising of lights' or 'surfeit', however, produces a match using the ICD10 online tool.

Each preliminary code from Checks 2 and 3 should also be checked against the Masterlist in case there is a more specific ICD10h code. Using the examples above, in the case of 'cholera asiatica' HistoricStringsEnglish yielded A00.900. The Masterlist indicates that this is a code for 'Cholera, unspecified' and there are no ICD10h codes for specific forms of cholera. A00.900 is therefore the final ICD10h code. In the case of 'aplastic anaemia' there are no specific historic codes, so the final ICD10h code is D61.900. In the case of 'green sickness' however, we see that a specific historic code, D50.801 for 'chlorosis', has been created under D50.8. We then need to consider whether to code 'green sickness' as D50.801 or as the more general D50.800. In other words, were 'green sickness' and 'chlorosis' the same condition? A search of the historic literature and description of the symptoms of each suggests that they were the same, therefore the final code for 'green sickness' should be D50.801.

Finally we can return to the strings not found in either HistoricStrings, the Masterlist or ICD10, such as the examples used above of 'rising of lights' and 'surfeit'. In such instances we need to consult the historic literature about the meaning of the terms. 'Surfeit' was used for general over-indulgence, and can therefore be coded to R63.200: 'polyphagia' or excessive eating. 'Rising of lights' is more difficult. 'Lights' is an old word for lungs but it is unclear what the 'rising' referred to. In the absence of more conclusive evidence, R09.800 'Other specified symptoms and signs involving the circulatory and respiratory systems'

[23] Cholera asiatica, or more commonly asiatic cholera, was a term commonly used for the second cholera pandemic, 1826-1837.

seems the best code. If there are many deaths from the cause a new code could be requested (see section 2.4.ix).[24]

Further examples are explained below and all the examples used are summarised in Table 3.

'Spasm'

Searching HistoricStringsEnglish comes up with various CODs including 'spasm', from 'spasmodic diarrhoea', heart spasms, stomach spasms, bowel spasms etc. The last is just 'spasms' R25.200. If your cause is simply 'spasm' then this is the correct choice. NB A search in the Masterlist also finds stomach spasm, bowel spasm etc. and reveals that R25.200 is 'cramp and spasm'. Searching the ICD10 online tool comes up with a longer list which include 'spasm', again 'cramp and spasm' R25.2 is revealed as the right choice as it is the option without any further explanation. However, if you have identified a code using the ICD10 resource it is important to go back to the codes under R25.2 in ICD10h to see if a specific historic code has been created. Remember that we code the word, not the meaning, so don't worry if you think that 'spasm' is less of a muscle twitch and more like a convulsion – we deal with that by grouping them together in the classification step.

'St Anthony's fire'

HistoricStringsEnglish reveals that this is an archaic word for erysipelas, and should be coded to A46.001. The Masterlist gives the same answer. The online version of ICD10 also directs you to Erysipelas (A46) (but beware that MESH directs you to ergotism T62.2 – see list of ambiguous causes). If you have found the code A46 in the ICD10 online resource it is then important to go to the codes under A46 in the Masterlist, where you can find the specific code for St Anthony's fire.

'Flux'

HistoricStringsEnglish and the Masterlist indicate a code A09.960, under (presumed) infective diarrhoea. The online version of ICD10 is no help in this case as 'flux' is not a 21st diagnosis.

'Legionnaires disease'

HistoricStringsEnglish does not reveal any matches (this cause was not recognised until 1976 and so did not appear in the datasets used to create the system). However, the online version of ICD10 indicates that in ICD10 2016 this condition is given the code A48.1, and this is also in the MasterList which contains all the original ICD10 codes and their descriptors. It is still important to check the ICD10h codes associated with A48.1 to see if any historic versions have been added which might be suitable. In this case they have not, and therefore the ICD10h version of the ICD10 code, A48.100, should be assigned.

[24] Rising of lights was a cause which occurred regularly in the 17th and 18th century in the UK (there were 133 deaths from the cause in London in 1721, amounting to 0.5 per cent of the total). It does not occur in ICD10h because it was no longer in common usage by the mid-nineteenth century.

Figure 6

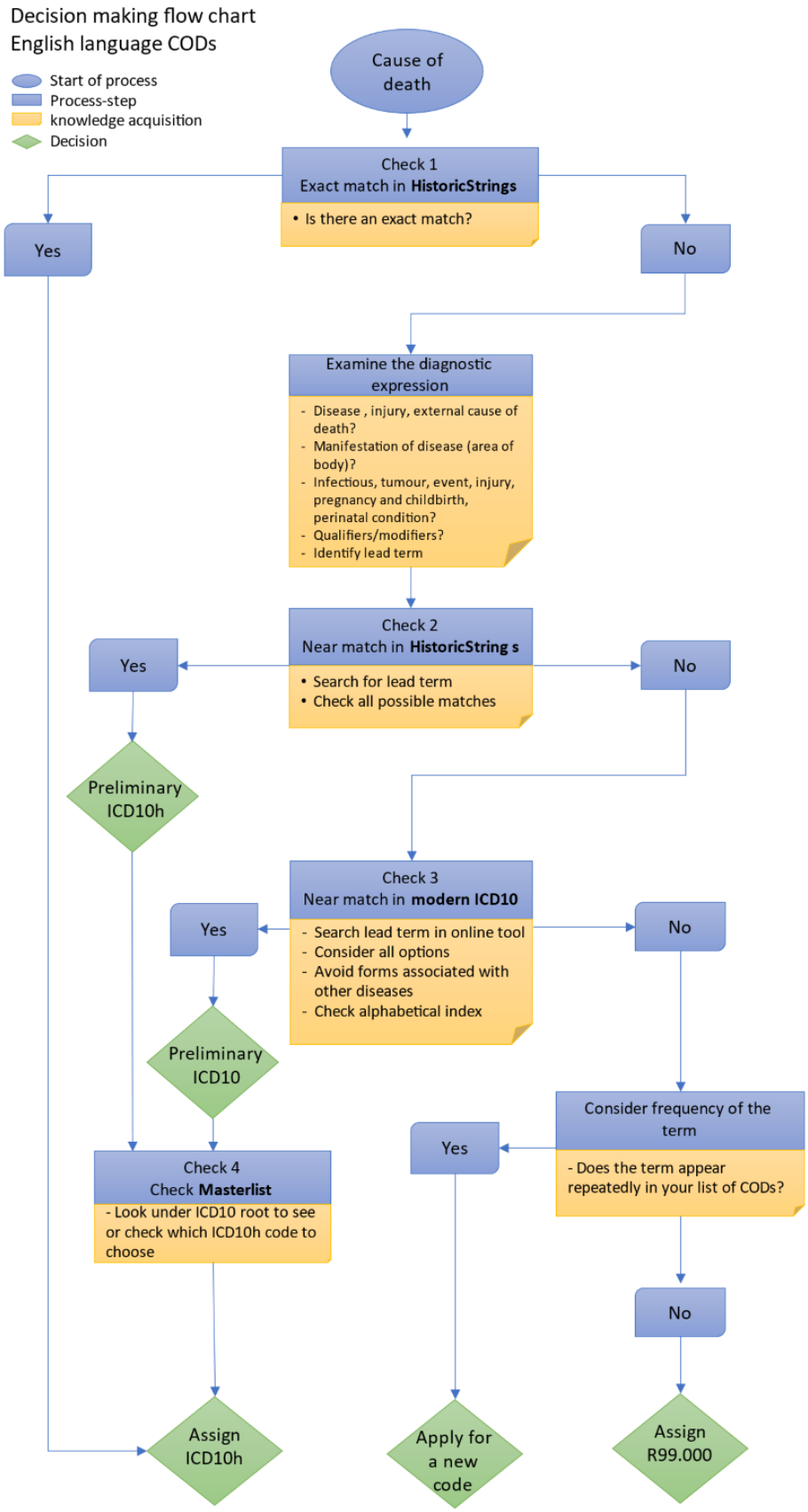

Decision making flow chart
English language CODs
Start of process
Process-step
knowledge acquisition
Decision
Cause of death
Check 1
Exact match in HistoricStrings
• Is there an exact match?
Yes
No
Examine the diagnostic expression
- Disease , injury, external cause of death?
- Manifestation of disease (area of body)?
- Infectious, tumour, event, injury, pregnancy and childbirth, perinatal condition?
- Qualifiers/modifiers?
- Identify lead term
Check 2
Near match in HistoricString s
• Search for lead term
• Check all possible matches
Yes
No
Preliminary ICD10h
Check 3
Near match in modern ICD10
- Search lead term in online tool
- Consider all options
- Avoid forms associated with other diseases
- Check alphabetical index
Yes
No
Preliminary ICD10
Consider frequency of the term
- Does the term appear repeatedly in your list of CODs?
Yes
No
Check 4
Check Masterlist
- Look under ICD10 root to see or check which ICD10h code to choose
Assign ICD10h
Apply for a new code
Assign R99.000

Table 3: English language coding examples

Original string	HistoricStrings English	ICD10h	ICD10h description	ICD10	ICD10 description
Cholera asiatica	Asiatic.cholera	A00.900	Cholera, unspecified	A00.9	Cholera, unspecified
Aplastic anaemia	-	D61.900	Aplastic anaemia, unspecified	D61.9	Aplastic anaemia, unspecified
Green Sickness	-	D50.801	Chlorosis	D50.80	Other iron deficiency anaemias
Surfeit	-	R63.200	Polyphagia	R63.2	Polyphagia
Rising of lights	-	R09.800	Other specified symptoms and signs involving the circulatory and respiratory systems	R09.8	Other specified symptoms and signs involving the circulatory and respiratory systems
Spasm	Spasms	R25.200	Cramp and spasm	R25.2	Cramp and spasm
St Anthony's fire	St.Anthony's. fire	A46.001	St Anthony's fire	A46	Erysipelas
Flux	Flux	A09.960	purging/flux	-	-
Legionnaires disease	-	A48.100	Legionnaires disease	A48.1	Legionnaires disease

2.4.v Conditional codes

There are some causes which have different codes according to the age or sex of the deceased. When coding large volumes of causes and assigning codes to unique COD strings we cannot take this into account. The proposed solution, if codes have been assigned to individual deaths, is to give a default code and then check codes known to be affected by this issue afterwards, reassigning the codes to reflect the person's sex and/or age at that point. Alternatively, the categorisation of particular codes can be adjusted to reflect particular age-sex groups. Examples include:

- Childbirth – if it is the mother who dies this should be coded under codes starting O (e.g. O95.000 for childbirth/obstetric death of unspecified cause) but if it is the child who died it should be coded under P (e.g. P03.900 for newborn affected by complication of labour and delivery, unspecified). The ICD10h default here is to assume it is the mother who dies. Any O codes assigned to infants should be reassigned to a suitable P code in the later stage. O and P codes are directed to different classes in HistCat, so it can be important to perform this reallocation, or at least assess the effect of not doing so.
- Premature birth – again this could refer to either the child or the mother, and the default is to assign the child code (P07.300) and reassign when the causes have been added to the individual deaths.
- Bronchitis – chronic and acute bronchitis are given different codes in ICD10. Where acute or chronic is specified, allocating a code is straightforward. However, where it is not specified, ICD10 places cases of bronchitis of <4 weeks duration or suffered by those under age 15 with acute

bronchitis; all other cases are assumed to be chronic. The ICD10h default is therefore under the chronic codes, with the specific code J40.009 (bronchitis, unspecified as to age, duration or acute or chronic). In data sets where age and or duration of illness are known these codes can be reallocated to an 'acute' code at a later stage.

- Slag or slaganfall - a Swedish and Norwegian term for stroke, but in certain historical periods often used for sudden death in infancy. The default code is a specific code for slag among adults, I64.003, which allocates deaths from this cause to stroke. ICD10h contains a separate code, R95.901, for slag (or slaganfall) among infants and small children, and if there are many deaths from slag among infants and children these can be reassigned to R95.901.
- Various codes referring specific male or female reproductive organs, particularly (but not limited to) in cancers and genitourinary diseases. A full list of sex-specific codes is provided in Annex 7.8 (pp 227-234) of the ICD10 coding manual, and these are also identified in the 'GenderSpecific' column in the Masterlist (where 1 identifies diseases which should only be used for men, 2 indicates diseases which should only be used for women, and 0 identifies diseases which can be used for both men and women).

In general we recommend checking the age and sex distribution of the HistCat distribution after codes have been assigned to deaths. Deaths categorised to 'childbirth' and not occurring among women aged 15 to 49 should be checked, as should those categorised to 'perinatal' and occurring to those over age one and those categorised to 'old age' and occurring to those under age 50.

2.4.vi Cancers/Neoplasms

As explained above, 'cancers' or 'neoplasms' is the one causal group which is treated slightly differently in ICD10h compared to ICD10, and we code all pre-twentieth century cancers to chapter C, rather than assigning those without histologically confirmed malignancy to Chapter D. Anything which suggests cancer should therefore be placed in the relevant place in the C chapter, and care should be taken to use the applicable code for one of the main terms for cancer: cancer, scirrhus, sarcoma, epithelioma, tumour, malignant disease, and carcinoma, which receive codes .x01 to .x07 respectively, as per Table 4. The more modern term 'neoplasm' is given code .x00.

For example, in ICD10 a cause which stated 'tumour of the breast' should be coded to D48.6: breast neoplasm of uncertain or unknown behaviour, unspecified. Breast cancer (confirmed malignant) should be coded to C50.9: malignant neoplasm of the breast, unspecified. Instead of following this, we assume that anything containing any of the words cancer, scirrhus, sarcoma, epithelioma, tumour, malignant disease, and carcinoma denotes malignant neoplasm and therefore we can search ICD10 for malignant neoplasm of the breast. As usual, we should then return to the codes under C50.9 in the Masterlist to find the specific historical code for 'tumour of the breast', which is C50.905. Similarly, the historical code for 'breast cancer' is C50.901. Those which are **specified** as being benign (e.g. 'benign tumour of the breast') should be coded to codes starting with D (in this case D48.600), as per ICD10. Those which are specified as being of uncertain behaviour (e.g. 'breast cancer - uncertain behaviour') should also be given a code starting with D (in this case D48.600).

Table 4: Examples of coding breast cancer (part of breast unspecified)

Historic string	ICD10h	ICD10h_description
malignant neoplasm of breast	C50.900	malignant neoplasm, breast unspecified
breast cancer	C50.901	cancer, breast unspecified
scirrhus of breast	C50.902	scirrhus, breast unspecified
sarcoma in breast	C50.903	sarcoma, breast unspecified
epithelioma in breast	C50.904	epithelioma, breast unspecified
tumour, breast	C50.905	tumour, breast unspecified
malignant disease of breast	C50.906	malignant disease, breast unspecified
breast carcinoma	C50.907	carcinoma, breast unspecified
lymphoma breast	C50.908	other form of cancer, breast unspecified
benign tumour the breast	D48.600	Neoplasm of uncertain or unknown behaviour of other and unspecified sites|Breast

In order to preserve comparability with ICD10 which separates cancers/neoplasms with confirmed malignancy from those without, researchers may allocate a code in a separate field which indicates whether the string identified the neoplasm as being malignant, benign, stated to be uncertain behaviour, or if there was no information on malignancy. This indicator, here called CancerMalignancyFlag, should be applied to all codes in the ICD10 neoplasms chapter/HistCat neoplasms category (codes C00 to C97 and D00 to D48), and can be identified by the presence/absence of qualifying words in the historic strings. Table 5 provides the values of the CancerMalignancyFlag and their interpretation, and the HistoricStringsEnglish file provides examples of how to use these in a column of the same name.

Table 5: Cancer malignancy flag values and interpretation

Flag value	Interpretation
0	Not a cancer/neoplasm
1	No information on malignancy
2	Stated to be malignant
3	Stated to be benign
4	Stated to be of uncertain behaviour

2.4.vii External causes (including accidents, suicide and violence)

ICD10 has two sets of codes for accidents, violence and other 'external' causes, which under ICD10 rules should both be applied. One set is for the nature of the injury sustained ('injury code') and the other is for circumstances surrounding the injury, including motive or intent ('circumstance code'). Injury codes start with letters S and T, and circumstance codes start with letters V, W, X and Y. Therefore, suicide by hanging; suicide by gunshot; accidental gunshot; and murder by gunshot would all be given different code combinations. Historic causes often give a lot of information about either the nature of the injury or the

circumstances surrounding an ‘external’ cause of death, but do not always allow both injury and circumstance codes to be assigned.

As mentioned above, although more than one code can be assigned, as with ICD10, a circumstance code should always be given as this allows intent to be distinguished (where it is clear). Thus an individual who simply died of ‘drowning’ could be given the injury code T75.100, indicating simply ‘drowning and nonfatal submersion’ with no indication of whether the intent was known, but they should also be given code Y21.000, ‘drowning and submersion undetermined intent’, and if only one code is given it should be the latter. Similarly the death of someone who committed suicide by throwing themselves in a river could be coded both T75.100 and X71.000 (‘intentional self-harm by drowning and submersion’). The circumstance codes allow more options for categorisation: both the simple ‘drowning’ death and the ‘drowning by suicide’ death can be categorised as ‘drowned’, but the second could also be categorised with other deaths by suicide.

In the above example of drowning, allocation of the injury code does not add any more information to the circumstance code, and therefore there is no particular benefit to allocating it. Sometimes, however, the injury itself may be described in considerably more detail than is reflected in the circumstance codes. For example, there is a specific injury code for ‘crushed chest’, S28.000, but the circumstance code for crushed chest would be Y34.000 ‘unspecified event, undetermined intent’ which could also cover a wide variety of other injuries. In this case, therefore, the injury code can be used to add information. If the description of the injury also includes detail about how the injury came about (e.g. 'crushed chest in combine harvester accident’) a more informative circumstance code could also be given – in this case W30.000 denoting accidental death due to contact with agricultural machinery. In either case if just one code is allocated it should be the circumstance code. If more than one code is allocated, when a code representing the underlying cause of death is assigned, this should also be the circumstance code.

The complexity of external causes and the fact that a death from accident or violence may have several different parts means that it is often possible to separate each external cause into different fields. For example ‘crushed chest in combine harvester accident’ could be parsed into ‘crushed chest’ and ‘combine harvester accident’. If we are dealing with unique parsed causes, and encounter an injury such as ‘crushed chest’ we will not know whether the cause of death string from which it came also yielded another parsed string which can be given a more informative circumstance code, and in such cases we need to allocate both injury (‘crushed chest’) and circumstance (‘unspecified event, undetermined intent’) codes. However if we have sight of the original string, or of both parsed fields together, we will know that we can give an injury code to ‘crushed chest’ and a circumstance code to ‘combine harvester accident’. When the codes for parsed strings are reassigned to deaths, the most informative circumstance code can then be chosen.

In the ICD10h HistoricStringsEnglish table, when referring to external causes the ICD10h field always contains the circumstance code, with the ICD10hDescription field containing the corresponding description from the ICD10h Masterlist. The ICD10hInjury and ICD10hInjuryDescription fields contain injury codes and their descriptions where these are relevant. In a few cases the Injury fields can also contain other information, such as amputation (unspecified as to whether it was traumatic or surgical, assigned to codes under Z89).

As it can be difficult to retain the sense of the text strings related to external causes when parsing them, due to the fact that they often take narrative forms, it may be easier in the long-run to identify external causes before the parsing stage, and code the unparsed strings.

When coding deaths from external causes it can be helpful to start by considering the different 'blocks' of circumstance codes in ICD10 (see Table 6) and which, if any of them, best applies to the cause of death. Sections II (external causes of injuries) and III (drugs and chemicals) of the ICD10 Alphabetical Index can also be very helpful here.

Table 6: The division of ICD10 Chapter XX (external causes) into blocks, and their ICD10h code ranges

Block	**ICD10h codes**	**Details**
Accidents	V01.000-X59.900	Generally mention or imply an accident (include transport collisions and falls in places unlikely to be suicide).
Event of undetermined intent	X60.000-X84.000	Contain explicit words that state the intention – general strings such as 'drowning' 'gunshot wound' or 'poisoned' without further explanation should be placed here.
Intentional self-harm	X85.000-Y09.002	Include words such as 'suicide' or 'self-inflicted'.
Assault	Y10.000-Y34.007	Include words such as 'murder', 'manslaughter', 'assault' etc.
Legal intervention and operations of war	Y35.000-Y36.900	Mention 'war', 'legally executed' etc.
Complications of medical and surgical care	Y40.000-Y84.900	The cause of the care should also be coded (bear in mind it might be coded separately if that part of the original string was parsed into a separate field). The allocation of primary or underlying causes will be discussed in a subsequent section, but it is relevant to state that in these cases of complications of medical care, the primary cause should be allocated to the complication of the care only if the condition leading to the care was not likely to have resulted in a death in the absence of the care (e.g. infection following operation for ingrown toenail). Otherwise it should be allocated to the condition which led to the care.
Sequelae of external causes of morbidity and mortality	Y85.000-Y89.900	These codes should only be used in conjunction with a code regarding the injury or illness.
Supplementary factors related to causes of morbidity and mortality classified elsewhere	Y90.000-Y98.000	Contributory factors such as alcohol influence and lifestyle. These codes should not be used on their own. For example F10.000 (extreme intoxication) or F10.200 (alcoholism) should be used in preference to the codes for the specific degrees of intoxication and blood alcohol level indicated by codes in Y90 and Y91.

Some, but not all, of these groups have helpful subgroups. For example, the Accidents block (V01.000-X59.900) is separated into 'Transport accidents' (V01-V99) and 'Other external causes of accidental injury'

(W00-X59). These then contain further subgroups: 'Transport accidents' contains twelve subgroups relating to whether the accident was on land, water or in air, and for land transport, the position of the injured person (whether pedestrian or in a vehicle).

ICD10(2016) also provides the opportunity to add further codes as separate variables for the place that external events occurred (eg home, school, street etc.) and the activity of the injured person at the time of the event (eg working, sleeping, doing sports etc.). We have chosen not to incorporate these additional codes into ICD10h.

The tables below provide some examples of coding different strings relating to gunshots (Table 7) and drowning (Table 8). In the cases of gunshot wounds, the applicable injury codes rarely give much detail about where in the body the injury occurred, and there are injury codes which can provide more detail about this. In contrast, in the case of drowning, injury codes do not add more information, so there is no need to assign them.

Table 7: Examples of gunshot wounds, classification and ICD10h codes (examples from Sweden but translated into English)

Historic string	ICD10h - circumstance	Block - circumstance	ICD10h description - circumstance	ICD10h - injury	ICD10h description - injury
Accidently shot while moose hunting, gunshot wound	W34.000	Accidents	Discharge from other and unspecified firearms	T14.100	Open wound of unspecified body region
Gunshot injuries to the head. Revolver. Suicide.	X72.000	Intentional self-harm	Intentional self-harm by handgun discharge	S01.900	Open wound of head, part unspecified
Gunshot wound, suicide	X74.000	Intentional self-harm	Intentional self-harm by other and unspecified firearm discharge	T14.100	Open wound of unspecified body region, puncture wound with (penetrating) foreign body
Gunshot wounds to the head (murder)	X95.000	Assault	Assault by other and unspecified firearm discharge	S01.900	Open wound of head, part unspecified
Gunshot wound (homicide)	X95.000	Assault	Assault by other and unspecified firearm discharge	T14.100	Open wound of unspecified body region, puncture wound with (penetrating) foreign body
Gunshot wound in the right temple	Y24.000	Event of undetermined intent	Other and unspecified firearm discharge, undetermined intent	S01.800	Open wound of other parts of head
Gunshot wounds (in war)	Y36.400	Legal intervention and operations of war	War operations involving firearm discharge and other forms of conventional warfare	T14.100	Open wound of unspecified body region, puncture wound with (penetrating) foreign body

Table 8: Examples of drowning and submersion, classification and ICD10h codes (examples from Sweden but translated into English)

Historic string	ICD10h - circumstance	Block - circumstance	ICD10h description - circumstance	ICD10h - injury	ICD10h description - injury
Drowning while paddling in a canoe on the Lais River	V90.500	Accidents	Accident to watercraft causing drowning and submersion |Canoe or kayak		
Drowned by accident during timber floating	W69.000	Accidents	Accidental drowning and submersion |Drowning and submersion while in natural water		
The mother drowned the child, murder	X92.000	Assault	Assault |Assault by drowning and submersion		
Suicide by drowning	X71.000	Suicide	Intentional self-harm |Intentional self-harm by drowning and submersion		
Death by drowning	Y21.000	Event with undetermined intent	Event of undetermined intent |Drowning and submersion, undetermined intent		

2.4.viii Dagger and asterisk codes

In ICD10 (see section 3.1.3 in the ICD10 instruction manual) there are special pairs of codes called dagger and asterisk codes for causes of death which contain information about both an underlying general disease (the dagger code) and a manifestation in a particular organ or site which is a clinical problem in its own right (the asterisk code). For example in ICD10, 'Tuberculosis of the bladder' can be given two codes: the dagger code A18.1 indicating 'Tuberculosis of the genitourinary system', and the asterisk code N33.0 indicating 'Bladder disorders in diseases classified elsewhere'. Under mortality coding in ICD10 the dagger code should always be given but the asterisk is optional (it being of most use for morbidity coding). Asterisk codes should never be used alone. In ICD10h we recommend not using asterisk codes at all – indeed for common manifestations of particular diseases in different parts of the body ICD10h has created specific codes under the main disease. For example ICD10h has the code A18.102 for 'Tuberculosis of the bladder'.

In ICD10, causes where one disease is used as a modifier or adjective, such as 'tuberculous meningitis', could be given both the dagger code (in this case A17.0, indicating that the underlying disease is tuberculosis), and the asterisk code (G01.0, indicating that the lining of the brain is where the disease has

manifested). However, in ICD10h we recommend just giving the dagger code for tuberculous meningitis, which is A17.000. Note that if causes of death have been parsed out into separate fields, e.g. if tuberculous meningitis has been separated into tuberculosis and meningitis, they will be coded separately, and given codes A16.905 and G03.900. These can be resolved at a later stage (see section 2.5.ii), but for this reason it is preferable not to parse out phrases such as tubercular meningitis into tuberculosis and meningitis.

Although we recommend not using asterisk codes in ICD10h, we mention them here so that coders are alert to their presence (which will always mention 'in a disease classified elsewhere') and can ensure they do not use them in coding. For completeness, they are retained in the Masterlist, but are identified by a 1 in the 'DoNotUse' field.

2.4.ix Requesting new codes

It is tempting for coders to assign new codes where they think they need one, but we very strongly urge you not to do this, and instead to apply to the ICD10h Coding Committee if you would like a new code. This is because ICD10h is intended to be a standard international system, with comparability in codes across countries as well as across time. If different users assign different causes to the same code this will violate the principles of comparability and consistency and make it very difficult for an international database of historic causes to be generated from the language specific strings that individuals encounter in their research (see Section 3 for generating international lists).

Previous experience with members of the international SHIP network and the Digitising Scotland project who road-tested early versions of ICD10h found that most historic causes could be assigned an appropriate code from those in the Masterlist. If, however, you encounter a cause which does not have a good fit in the Masterlist, first ask yourself whether it is a common cause in your data. If not, then it is best to assign it R99.000 for 'unclear'. If you still feel you need a new code, then please contact the coding committee (please see GreatLeap.eu for details of how to do this), with your cause in the original language, an English translation, a suggested code, and a justification of why the code is needed.

2.5 Apply codes to individual deaths; assigning primary causes and classification

2.5.i Applying codes for unique strings to the entire dataset

When you have finished coding each unique string, use the look-up tables you created in your data preparation stage (see Section 2.3) to assign the ICD10h code to the first, second, third etc. causes for each individual. If an individual has more than one cause of death, they will end up with more than one code. External causes are also likely to end up with more than one code.

At this stage it is sensible to reallocate certain ambiguous codes (particularly childbirth) according to age and sex. Other ambiguous causes (e.g. bronchitis) will probably all end up in the same classification so this is not so important, unless you want to look at the development of causes within the general classes of HistCat (or another classification).

2.5.ii Dealing with multiple causes and assigning an underlying cause of death

Deaths with more than one cause can be particularly interesting and informative, particularly when looking at combinations of conditions and changes in the assignation of causes of death over time. They can allow causes which are often complications of some other cause to be examined more fully, for example pneumonia, convulsions and so on. However, they are also complicated to deal with, particularly when trying to decide what 'overall' or 'underlying' cause the deceased died from. Most researchers want to assign such an overall cause in order to look at the distribution of deaths from different types of disease.

ICD10 defines an underlying cause as that which might have started a process leading to death ('the disease or injury which initiated the train of morbid events leading directly to death, or … the

circumstances of the accident or violence which produced the fatal injury' (ICD10 manual p.31). Therefore, if both 'measles' and 'pneumonia' appear on the death certificate the most plausible chain of events is that the deceased caught measles and then developed pneumonia as a complication, so measles should be the underlying cause, irrespective of the order in which they appear on the certificate. A full set of instructions for assigning the underlying cause of death, with examples, can be found in the ICD10 manual pp. 31-139. This also includes lists of causes which should *not* be used as underlying causes, including most vague and ill-specified causes of death. In a historic context we try to follow the rules but it is not always possible to avoid a vague underlying cause of death when all the causes listed are vague.

There are programmes which can automatically assign the underlying cause (e.g. https://www.cdc.gov/nchs/nvss/mmds/about_mmds.htm) but they tend not to work very well for historic causes. The ICD10h team is working out a programmatic solution to helping assign the underlying cause of death for historic data according to the ICD10 guidelines as applied to ICD10h codes and taking historical idiosyncrasies into account.

Many researchers, however, take the easier option of assuming the death certification instructions regarding the ordering of underlying and other causes on the death certificate in their country were followed by doctors. For example, in the UK in the 19th century, doctors were supposed to place the underlying cause first on the certificate, and this should mean that taking the first cause will capture most underlying causes. Nevertheless, inspection of the individual certificates indicates that doctors did not always follow this advice. Some researchers therefore take the first cause, unless it is vague or ill-defined, in which case they move onto the next cause on the same certificate to see if there is one which is not ill-defined. However, it should be noted that this does not solve issues related, for example, to doctors writing 'pneumonia following measles' instead of 'measles followed by pneumonia' and this can result in systematic distortions. Researchers should always be careful to state which process they used to assign the overall causes of death when reporting their research.

2.5.iii Categorising deaths

Once one or more codes have been assigned to each death, it is easy to also apply the HistCat categorisation which is provided in the Masterlist, or one of the InfantCat categorisations which are provided separately (InfantCat was used for the comparative exercise using ICD10h published in the contributions to the special issue 'What was killing babies in port cities?' in the Journal of Historical Life Course Studies, and InfantCat2024 is a revised infant classification incorporating suggestions arising from this comparative exercise). See Table 9 for details of these. More categorisations are being produced, including a categorisation for child mortality. We also intend to produce categorisations with specific adjustments for certain countries to reflect the different meanings of particular terms.

N.B. A very small numbers of ICD10h codes have 'not a cause of death' in the HistCat and InfantCat columns. These codes should not be used as underlying causes of death, so they should not appear as a category. If one of these codes is the only code for a death, that death should be allocated to the 'No cause given' category.

Table 9: Historical categorisations

Categorisation	version	description/adaptations	date
HistCat, General categorisation designed to enable comparison with published Scottish nosologies 1855-1950	HistCat	Childbirth Circulatory Debility* Diarrhoea Digestive Genitourinary Ill defined Infectious Neoplasms Nervous system no cause given/blank Old age Other Perinatal Respiratory stated to be 'unknown' Tuberculosis External causes**	October 2020
InfantCat, specifically for infant deaths	InfantCat2020	Airborne diseases Food and water borne diseases Other infectious diseases Congenital and birth disorders Weakness Convulsions Teething External causes Other Ill-defined Stated to be unknown No cause given	October 2020
	InfantCat2024	Common infectious diseases Broncho-pneumonia Other (includes all causes from previous 'other', 'other infectious', and 'external' categories, and a few from 'airborne') Food-water borne Perinatal and weakness (combines 'congenital and birth disorders' and 'weakness' categories) Convulsions Ill-defined (combines 'ill-defined', 'stated to be unknown' and 'teething') No cause given	July 2024

*We recommend reallocating deaths in the debility category as follows (see Section 1.4 for background):

- If HistCat = Debility and age<1 then allocate to Perinatal
- If HistCat = Debility and age >=70 then allocate to Old Age
- Otherwise if HistCat=Debility allocated to Ill-defined

**In the version of HistCat published in 2020 this category was called 'Violence'.

2.6 Updated versions of ICD10h

This release of ICD10h (2024) is based on a prototype version circulated among members of the SHIP project in 2020. A look-up table has been produced between the two versions, indicating the new (2024) ICD10h code to which each 2020 ICD10h code corresponds. Similar tables will be produced for each revision of the system.

3. Citing ICD10h

ICD10h is a complex and evolving system, and will hopefully end up having many contributors. In order that every contributor's work is properly acknowledged, we ask that you pay close attention to the following citation guidelines, even though they are rather complex.

Please cite every element of the ICD10h system that you use, not just the MasterList or the Manual, and make sure that the version and DOI are correct.

Each of the following has a separate citation and DOI:

ICD10h_Manual (this document)

- Alice Reid, Eilidh Garrett, Maria Hiltunen Maltesdotter, Mayra Murkens, 2024, ICD10h: Historic cause of death coding and classification scheme for individual-level causes of death - Manual [https://doi.org/10.17863/CAM.109960]

ICD10h_Masterlist_2024

- Alice Reid, Eilidh Garrett, Maria Hiltunen Maltesdotter, Angelique Janssens, 2024, ICD10h: Historic cause of death coding and classification scheme for individual-level causes of death - Codes [https://doi.org/10.17863/CAM.109961]

ICD10h_HistoricStringsEnglish

- Alice Reid, Eilidh Garrett, Maria Hiltunen Maltesdotter, 2024, ICD10h: Historic cause of death coding and classification scheme for individual-level causes of death – English language historic strings [https://doi.org/10.17863/CAM.109962]

ICD10h_InfantCat

- Alice Reid, Eilidh Garrett, Maria Hiltunen Maltesdotter, Angelique Janssens, 2024, ICD10h: Historic cause of death coding and classification scheme for individual-level causes of death – Infant Categorisations [https://doi.org/10.17863/CAM.109963]

Additional HistoricStrings in other languages and new categorisations produced in between revisions of ICD10h will be identified on the greatleap.eu website with their own DOI and added in the next revision of this document.

4. Contributing to the international version of ICD10h

Following the above steps for allocating codes can be extremely time consuming, particularly when it comes to non-English language causes. Data-sets in the English language can be coded much more quickly due to the existence of the HistoricStringsEnglish file which lists the strings found in some historic English language datasets. The production of similar lists for other languages would make coding new data in those languages much easier. We therefore encourage those who have coded their data to ICD10h to submit lists of strings and ICD10h codes to the ICD10h database (please consult the GreatLeap.eu website for

instructions on how to do this). It is very important, however, that such datasets have been coded carefully and in accordance with the instructions set out above, and that no new codes have been assigned without prior approval of the ICD10h coding committee. This will help maintain consistency and comparability across different languages.

We also realise that some countries may need to adapt HistCat and other categorisations to adjust for different common historical uses (for example 'angina NOS' is coded into diseases of the circulatory system in HistCat, but in the Netherlands it was most commonly used for a form of tonsillitis, so the Netherlands version moves it to Respiratory disease). We would also like to collect country-specific versions of HistCat, clearly documented with differences to the regular HistCat and justification of those differences.
In addition, a working group will be set up to produce mappings of ICD10h codes onto other historic categorisations, such as other versions of the ICD system, the Bertillon system and so on. Any such mappings which have already been produced will be gratefully received and presented with their original authorship and separate DOI. Please see the GreatLeap.eu website for details on how to contribute.

5. Appendices

5.1 **Ambiguous cases and 'false friends'** (please let us know of others you come across)

Angina: In the English language, angina is most often used to describe the heart affliction 'angina pectoris' (I02.900). However the word angina is also part of other, quite different, disease terms, such as Ludwig's angina (K12.2 a form of mouth abscess), Vincent's angina (A69.1, a form of pharyngitis or tonsillitis) and Diphtheritic membranous angina (A36.0, a form of diphtheria). 'Angina' with no further specification should be coded to I02.901 and can then, if the historic evidence suggests the predominant pre-twentieth century usage was different, the country-specific HistCat can be adjusted (as in the Dutch case).

Childbirth/difficult labour and certain other deaths during the birthing process could apply to either the mother (placed in the O chapter) or the infant (placed in the P chapter). The default cause normally is the O code (but this should always be checked), but it is highly recommended that the this should be reviewed once codes have been assigned to individual deaths and reassigned on the basis of age.

Chronic pneumonia in Swedish means Tuberculosis and is re-classified as such in the Swedish version of HistCat.

Cramps is a direct translation in many languages for convulsions, but there is also an English cause 'cramps' which tends to be milder, e.g. period pains. The meaning of the term in the modern version of ICD10 in the original language should always be followed.

Eclampsia in English always refers to puerperal eclampsia (15.900), but in other languages (e.g. Spanish and Dutch) it can be used to refer to convulsions or fits for other reasons, and the non-specific use of the term in the modern version of ICD10 in such languages should always be followed, which might entail them being coded with convulsions (R56.800).

Lung inflammation in some languages (e.g. Swedish and Dutch) means pneumonia and, in such languages, should be coded to J18.900. In English it is a much less specific term and should be coded to R09.101 where it has a specific historic code.

Nervous fever is generally said to refer to typhus and has been coded A75.901. However typhus and typhoid were often confused and sometimes categorised together before the twentieth century. In some countries 'nervous fever' was more often used to refer to typhoid fever after typhus and typhoid were more routinely distinguished. The specific code allocated for nervous fever will allow reclassification in country-specific cases where necessary.

St Anthony's Fire: this term has been used to refer to three different historic diseases – erysipelas, herpes zoster, and ergotism. We have coded it to erysipelas (A46.001).

Slag – a word which translates to 'stroke' in various Germanic languages and has therefore been allocated a specific code under stroke, I64.003. In English stroke always refers to a cerebrovascular incident (I64.000) but 'slag' can be used more broadly for convulsions and in some countries and time periods a considerable number occur among children.

Teething: often thought to indicate diarrhoea. We have coded teething and dentition to a variety of historic codes under the ICD10 code for teething syndrome (K00.7).

Typhoid: typhus (A75.900) and typhoid (A01.000) were often confused and sometimes categorised together before the twentieth century. See also nervous fever (A75.901).

Typhus: typhus (A75.900) and typhoid (A01.000) were often confused and sometimes categorised together before the twentieth century. See also nervous fever (A75.901).

5.2 ICD10h tables – full descriptions

Masterlist

IDMasterlist	unique ID number for Masterlist table
ICD10h	ICD10h code
ICD10	ICD10 code
ICD10_2levelCATEGORY	ICD10 first part of 2 level categorisation
ICD10_2levelCAUSE	ICD10 second part of 2 level categorisation
ICD10h_DESCRIPTION	ICD10h description (differs from ICD10_2levelCAUSE only where there is a specific historical code)
HistCat	HistCat category
DoNotUse	1=do not use for mortality coding (asterisk codes)
NotForUnderlying	1=do not use for underlying mortality codes
GenderSpecific	0=can be used for men or women; 1=use for men only; 2=use for women only

InfantCat

IDMasterlist	unique ID number, same as Masterlist table
ICD10h	ICD10h code
Infantcat2024	Infantcat2024 category
Infantcat2020	Infantcat2020 category

HistoricStringsEnglish

ID_HistoricStrings	unique ID for HistoricStringsEnglish table
HistoricString	Historic cause of death string as encountered in records
ICD10h	ICD10h code
ICD10hDescription	ICD10h description
ICD10hInjury	additional ICD10h code for injuries
ICD10hInjuryDescription	additional ICD10h description for injury code
CancerMalignancyFlag	0=not a neoplasm; 1=no information on malignancy; 2=stated to be malignant; 3=stated to be benign; 4=stated to be of uncertain behaviour

2020to2024transfer (published with the Masterlist)

ID2024_transfer	unique ID for 2020to2024transfer table
IDoct2020Masterlist	ID variable from the 2020 Masterlist
ICD10h_oct2020	ICD10h from the October 2020 Masterlist
ICD10h2024	ICD10h 2024 value

Figure 7: Relationships between ICD10h tables