SwedPop Documentation

Principles of Coding Historic Causes of Death Using the ICD10h Code System

Authors: Maria Hiltunen Maltesdotter and Annika Westberg, CEDAR

Version: February 26, 2026

Introduction

One of the primary objectives of the national infrastructure SwedPop is to provide researchers with harmonized and linked historical population data derived from Sweden's major historical databases. To ensure consistency and comparability across these datasets, SwedPop has implemented a systematic harmonization of cause of death information recorded in the original sources. This work has been undertaken by a specialized team of research analysts and system developers at CEDAR, Umeå University.

As part of this harmonization effort, SwedPop has adopted the ICD10h classification system—an international coding framework for historical causes of death developed through European research collaborations (Reid et al., 2024a). A comprehensive manual outlining the structure and application of ICD10h has been published by Reid et al. (2024b).

The purpose of the present document is to provide a technical overview of the methodology used to harmonize cause of death data specifically for Swedish historical sources, including death and burial registers and death certificates. It outlines the procedures employed to adapt these data to the ICD10h classification within the SwedPop infrastructure.

A cause of death can be defined as the chain of events leading to death. Causes of death may include diseases, accidents and other external causes of death such as injuries, poisoning, suicide and homicide. In historical Swedish sources, the causes of death information was recorded in the death and burial books and in death certificates. For further details on the historical reporting causes of death in Sweden, see Rogers (1999) and the National Board of Health and Welfare (2010).

Mapping historical causes of death to a modern coding framework presents significant challenges. One difficulty is that the cause of death information provided in the historical sources often described symptoms rather than explicitly identifying a disease. Additionally, from the nineteenth century onwards, there has been a major change in both medical terminology and diagnostics, resulting in temporal variation in cause of death data. Examples include the existence of multiple names for the same disease as well as the use of archaic disease names that are no longer used in modern medical terminology. Variations in who reported the death and, in particular, whether a physician was involved further contributes to inconsistencies in the historical records. Consequently, systematically interpreting and translating historical disease names into modern classifications is a complex and often ambiguous task.

The process of harmonization

The original data spans three centuries of historical sources, and its harmonization requires a series of crucial decisions before the information contained in the original text string is encoded. Interpretation challenges arise from factors such as illegible handwriting as well as the presence of abbreviations and semantic ambiguities that demand careful consideration.

Causes of death are distinguished between diseases and external events that led to death. External causes of death are handled in a separate process and are further explained below.

Splitting and standardizing causes of death

In the harmonization process, coding is performed using a single, standardized cause of death. Therefore, the first step involves splitting original text strings that contain multiple causes into individual components while keeping the order in which causes of death appear in the text string intact. This step is carried out automatically using a predefined set of rules, followed by manual validation and correction to ensure accuracy. Standardization is a wide concept, and, in this context, we mean standardizing different spellings into one and ordering the sequence that the cause of death appears in the text string. We also standardize the way in which the causes of death are separated when there is more than one cause of death mentioned in the string.

Principles of standardization

The primary objective of standardization is to preserve as much of the original information as possible while eliminating spelling inconsistencies and errors. Swedish historical sources contain numerous causes of death recorded in Latin; these are retained in Latin, but presented in a standardized format.

The standardization principles include:

1. **Normalization of spellings and abbreviations**

   The process aims to maintain the integrity of the original information while harmonizing spelling variations and abbreviations. For example, the Swedish terms for tuberculosis—*Tuberculos*, *Tuberkulos*, and *TBC*—are standardized to *Tuberkulos* in Swedish, whereas the Latin term *Tuberculosis* remains standardized as *Tuberculosis*. Similarly, Latin variants such as *Haemorr*. and *Haemoragia* are consolidated under the standardized form *Haemorrhagia*.

2. **Standardization without semantic interpretation**

   Causes of death are standardized without translating or modernizing archaic terminology. Even when the contemporary equivalent is known, the original term is preserved in its historical form. For instance, *Lungsot* (pulmonary consumption) is not converted to modern term is *tuberkulos i lungorna* (tuberculosis of the lungs).

3. **Separation of descriptive characteristics**

   Original entries often include descriptive characters such as *acute, chronic, epidemic, congenital, traumatic*, alongside the disease or injury and the affected organ. These characteristics are retained but placed in brackets following the standardized cause of death.

4. **Extension of contextual information**

   In certain cases, contextual details from the original text string must be manually extended to other causes of death within the same entry to ensure accurate coding, see Table 1 for examples.

Table 1. Example of extended separation and contextual standardization

Original text string **(in english for clarification)**	**Extended separation and standardization** **(in english for clarification)**
Childbirth, heart paralysis	Separated into: 1 Childbirth 2 Heart paralysis (at childbirth) Heart paralysis occurring during childbirth is assigned a specific code (O99.400) to indicate obstetric context, whereas other cases of heart paralysis are coded as I46.900. The childbirth context is explicitly used to convey that the condition manifested in connection with delivery.
Congenital weakness, heart failure	Separated into: 1 Weakness (congenital) 2 Heart failure (congenital) The expression *congenital weakness* is typically used in reference to newborns. Conditions associated with newborns are classified under a separate chapter in ICD10. In such cases, extended separation and standardization are applied to ensure that the attribute congenital is explicitly retained and incorporated into the standardized representation, as this information is essential for accurate coding.

Coding framework and adaption

The ICD10h system is based on the modern international classification ICD-10 (2016 edition), which has been adapted to accommodate medical terminology and designations prevalent in the 19th and early 20th centuries. For details on the system's construction, see Janssens (2021) and Reid et al. (2024a), Reid et al. (2024b). For an introduction to ICD-10, consult the ICD-10 Instruction Manual (WHO, 2016).

Using ICD-10 as a foundation, historical cause of death data has been organized into a structure tailored to historical disease nomenclature. It is important to note that strict adherence to all ICD-10 rules and guidelines is not feasible when coding historical causes of death. Instead, ICD10h applies the core principles of ICD-10 coding as a starting point and extends them to address the often vague, symptomatic, and context-dependent terminology found in historical sources. This approach necessarily involves interpretation and approximation during code assignment.

Standardizing and coding ambiguous text strings

While most cause of death text strings can be assigned standardized names that accurately reflect their content, some entries are too ambiguous or unclear to interpret with certainty. These texts may contain partial references to diseases or injuries but lack sufficient detail for precise classification. To ensure that potentially valuable information is not lost, these ambiguous entries are also assigned standardized names and corresponding codes, as illustrated in Table 2.

Table 2. Standardizing and coding ambiguous text strings

ICD10h	Standard name	Explanation
R69.000	EJ TOLKBAR SJUKDOM (Disease not possible to interpret)	The text contains information that with high probability can be interpreted as a description of a disease, but it is difficult or impossible to determine which disease it is.
R99.000	EJ TOLKBAR DÖDSORSAK (Cause of death not possible to interpret)	The text contains information that with high probability can be interpreted as a description of a cause of death, but it is difficult or impossible to determine which cause of death it is.
R99.099	EJ MÖJLIG ATT TOLKA (Not possible to interpret)	The text only contains information that is impossible to interpret. It is difficult or impossible to determine if the text describes a cause of death or something else.
R99.099	TEXTSTRÄNGEN ÄR INTE EN DÖDSORSAK (Text string is not a cause of death)	The text only contains information that is not a cause of death, for example a geographic location.

Principles of classification of external causes of death

Causes of death caused by external events are often complex, typically described as a sequence of events leading to injury and death. Such descriptions may include details about the circumstances and intent of the event, the nature of injuries sustained, the activity being performed, the location of occurrence, and, in some cases, consequential effects such as the onset of an illness that ultimately caused death. All these parameters must be considered when assigning ICD-10 or ICD10h codes for external causes.

Unlike diseases and injuries, external causes of death are not standardized in their original wording. They frequently appear as lengthy narratives, sometimes including names and geographic locations. Therefore, the

original text strings is first assigned an ICD10h-code. This code is then complemented by the ICD-10 SE description, which serves as the standardized name.

Table 3. Example of standardization of external causes of death; drowning by various intents

Original text string (in english for clarification)	ICD10h	Standard name (in english for clarification)
Fell overboard [*ship's name, date, location at event, travelling route*]. Body found at [*geographic location*]	Y21.000	Drowning and submersion, undetermined intent
Drowned in [*detailed description of location*], suicide under confusion of mind [*date of event and when found*]	X71.000	Intentional self-harm by drowning and submersion
Accidentally drowned in [*name of lake and location*] while swimming	W69.000	Drowning and submersion in sea, lake or watercourse
Murdered by its mother by drowning	X92.000	Assault by drowning and submersion

Final processing of standardized and coded causes of death

When a complete cause of death information for an individual is available, a set of rules is applied during the final processing stage.

1. Concatenation of identical ICD10h codes

Multiple causes of death sharing the ICD10h-code are concatenated in the standardized name field, separated by commas.

2. Additional coding for external causes of death

ICD10h applies two sets of codes for external causes, in accordance with ICD-10 rules:
- One for the circumstance and intent of the event
- Another for the injury sustained.

If injuries are recorded without details of the event, the entry is coded both for an event of undetermined intent and for the injury. For example, when the original text string only describes injuries, an additional code is assigned to represent the presumed circumstance. Both the standardized name and the ICD10h code are included in extracted datasets.

3. Contextual standardization for injuries

When information about the external event is available, the standardized name is derived from the ICD-10 description of the event. The injury is standardized and, where relevant, annotated with contextual information indicating that it resulted from a traumatic event. This distinction is critical when the injury overlaps with a disease category and requires different coding.

4. Substance-specific coding

Cases of poisoning by phosphorus are additionally coded with an injury code specifying the substance involved.

Table 4. Additional coding: External cause of death that lack circumstance of the event

Original text (not included in extracted data set)	**Processed and separated cause of death. Example of using context to extend standardized name (not included in the extracted data set)**	**Standardized name (included in extracted data set)**	**ICD10h (included in extracted data set)**
Combustio colli, brachii, thoracis et abdominis	Combustio abdominis, combustio brachii, combustio colli, combustio thoracis	Exposure to smoke, fire and flames, undetermined intent	Y26.000
Combustio colli, brachii, thoracis et abdominis	Combustio colli	Combustio colli	T20.000
Combustio colli, brachii, thoracis et abdominis	Combustio brachii	Combustio brachii	T23.000
Combustio colli, brachii, thoracis et abdominis	Combustio abdominis, combustio thoracis	Combustio abdominis, Combustio thoracis	T21.000

Table 5. Additional coding: External cause of death that include circumstance of the event

Original text (not included in extracted data set)	**Standard name (included in extracted data set)**	**ICD10h (included in extracted data set)**
Compressio medull. spinal (fall from railway vehicle)	Occupant of railway train or railway vehicle injured by fall from railway train or railway vehicle	V81.600
Compressio medull. spinal (fall from railway vehicle)	Compressio medullae spinalis (traumatica)	T09.300

References

Janssens, A. (2021). Constructing SHiP and an International Historical Coding System for Causes of Death. Historical Life Course Studies, 10, 64-70. https://doi.org/10.51964/hlcs9569

National Board of Health and Welfare (Socialstyrelsen) Befolkningsstatistik, Historik, produktionsmetoder och tillförlitlighet (2010) https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/statistik/2010-4-33.pdf

Reid, A., Garrett, E., Hiltunen Maltesdotter, M., & Janssens, A. (2025a). Historic cause of death coding and classification scheme for individual-level causes of death - Codes. Apollo - University of Cambridge Repository. https://doi.org/10.17863/CAM.109961.2

Reid, A., Garrett, E., Hiltunen Maltesdotter, M., & Murkens, M. (2025b). Historic cause of death coding and classification scheme for individual-level causes of death - Manual. Apollo - University of Cambridge Repository. https://doi.org/10.17863/CAM.109960.2

Rogers, J., (1999). Reporting causes of death in Sweden, 1750–1950. Journal of the history of medicine and allied sciences. 54(2): 190–209, https://doi.org/10.1093/jhmas/54.2.190.

WHO International Statistical Classification of Diseases and Related Health Problems 10th Revision (2016) https://icd.who.int/browse10/2016/en

The WHO ICD-10 Instruction Manual (2016) https://icd.who.int/browse10/Content/statichtml/ICD10Volume2_en_2016.pdf