Automated Coding of Historical Danish Cause of Death Data Using String Similarity

Perner, Louise Villefrance; Perner, Mads Linnet; Pedersen, Bjørn-Richard; Cañadas, Rafael Nozal; Sildnes, Anders; Shvetsov, Nikita; Andersen, Trygve; Bongo, Lars Ailo; Sommerseth, Hilde Leikny

Published in:
Digital Humanities in the Nordic and Baltic Countries Publications

DOI:
10.5617/dhnbpub.10662

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Perner, L. V., Perner, M. L., Pedersen, B.-R., Cañadas, R. N., Sildnes, A., Shvetsov, N., Andersen, T., Bongo, L. A., & Sommerseth, H. L. (2023). Automated Coding of Historical Danish Cause of Death Data Using String Similarity. In A. Rockenberger, S. Gilbert, & J. Tiemann (Eds.), *Digital Humanities in the Nordic and Baltic Countries Publications* (pp. 203-221) https://doi.org/10.5617/dhnbpub.10662

Automated Coding of Historical Danish Cause of Death Data Using String Similarity

Louise Ludvigsen[1], Mads Perner[1,2], Bjørn-Richard Pedersen[3], Rafael Nozal Cañadas[4], Anders Sildnes[4], Nikita Shvetsov[4], Trygve Andersen[3], Lars Ailo Bongo[4] and Hilde Leikny Sommerseth[3]

[1]*Saxo-Institute, Department of History, University of Copenhagen*
[2]*Section for Data Dissemination, Danish National Archives, Odense, Denmark*
[3]*Department of Archaeology, History, Religious Studies and Theology, UiT The Arctic University of Norway*
[4]*Department of Computer Science, UiT The Arctic University of Norway*

Abstract

The study of causes of death has been central to some of the most influential studies of the modern mortality decline in the nineteenth and twentieth centuries. The digitization of individual-level cause-of-death data has been game-changing, however, the data presents a major challenge: how do we code the thousands of unique strings for analysis in an efficient way? This paper aims to see how far we can get with automated coding based on string similarity. We do this by applying a Jaro Winkler string similarity algorithm in Python (pyjarowinkler) that codes our cause of death data from the Copenhagen Burial Register 1861-1911 to DK1875, a contemporary coding and classification system from nineteenth century Denmark. We then compare the performance of the algorithm to that of a manual (historian) coder in three different ways: at the level of each unique cause-of-death string, at the level of each cause-of-death group and for the overall cause-of-death pattern for all burials in Copenhagen 1861-1911. Our results show that a minimum-effort algorithm coded approximately half of the causes of death correctly compared to the manually coded dataset. This means that the method applied here is not accurate enough to use for actual data analysis of mortality patterns, as it is not possible to examine individual causes within larger causal groups. However, the results are promising for different uses of the method as a help for the manual coder. A way forward could be to use cut-off points of the Jaro-Winkler scores, coding only those causes where the string similarity match is relatively certain or use the automated method to catch most of the initial cases of a certain disease with a very set phrasing, such as cancer. In both cases, the remainder of the unique cause of death strings could then be coded by a manual coder.

Keywords

historical causes of death, string similarity, automated coding, mortality, individual-level data

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.
louise.ludvigsen@hum.ku.dk (L. Ludvigsen)
0000-0002-8708-4416 (L. Ludvigsen); 0000-0003-3890-207X (M. Perner); 0009-0000-2363-0791 (B. Pedersen); 0000-0001-6492-0218 (R. N. Cañadas); 0009-0003-0141-6112 (A. Sildnes); 0000-0002-8472-3702 (N. Shvetsov); 0009-0008-0893-367X (T. Andersen); 0000-0002-7544-2482 (L. A. Bongo); 0000-0001-7070-8184 (H. L. Sommerseth)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

1. Introduction

Historical causes of death are key to our understanding of one of the most revolutionary developments in population history: the great mortality decline of the nineteenth and twentieth centuries. Both Omran's theory of "The Epidemiologic Transition" [1] and McKeown's "The Modern Rise of Populations"[2] rely heavily on historical cause of death data in their attempts to describe the processes driving the mortality decline. One point of criticism has been their use of relatively sparse source material, consisting mainly of official aggregated statistics, published in the late nineteenth century and early twentieth century when these historical processes were still taking place [3, 4, 5, 6]. Contemporary statistics do indeed allow us to track mortality from certain diseases, but we are limited by the categories already defined by the physicians and statisticians who compiled the statistics at the time, according to their medical paradigms and knowledge. While some cause of death categories may be relatively straightforward to work with, such as "Mæslinger" ["Measles"], others hide a wide variety of diseases, such as "Andre sygdomme i ydre dele" ["Diseases in the outer parts of the body"][1]. The problem is especially acute when trying to study mortality over long periods of time or in different places, as parallel with actual changes in the disease landscape, the coding groups changed as the medical field developed. These difficulties are central in the discussion on how to work with historical causes of death, and what can be gained from doing so [8, 9, 10, 11, 12, 13].

Following years of work by archives, professional transcribers and the engaging community of citizen science and volunteers on digitizing historical sources, historical scholars now have access to unprecedented datasets and materials in both size and detail [2]. It has been a game-changer for long-term mortality studies that individual-level causes of death from historical sources such as parish registers, death certificates and burial records are now being digitized. With individual-level causes of death from the original handwritten records, we now have the opportunity for a much more in-depth analysis of the changes in disease patterns and in the factors that influence mortality risk in a given population and across generations.

In this paper, we make use of the Copenhagen Burial Register, which has been digitized and transcribed for the period 1861-1911 by the Copenhagen City Archives [15]. It contains the handwritten record of over 300,000 individual burials and more than 10,000 unique causes of death. This means that for the first time, it is possible to work with Danish individual-level causes of death from the nineteenth century at a large scale. However, the data presents a major challenge: how do we code the thousands of unique strings for analysis in an efficient way? Historians have traditionally preferred manual coding and consider this a gold standard. It is, however, a very time-consuming process, and since it also requires a considerable level of domain expertise, the method does not scale well and is unsustainable in the long run. One possible solution is to use machine learning algorithms, as has already been shown by several other projects on both historical and modern causes of death [16, 17, 18, 19]. However, machine learning is typically based on training data, which still relies on a time-consuming manual coding process by a person with large domain expertise. In addition, the earlier research

[1] The examples are taken from the Danish DK1875 system, used in the Danish aggregated cause of death statistics from 1876-1930. See appendix A and reference [7].

[2] An example of this is the SHiP collaboration working on individual-level causes of death in multiple European countries [14].

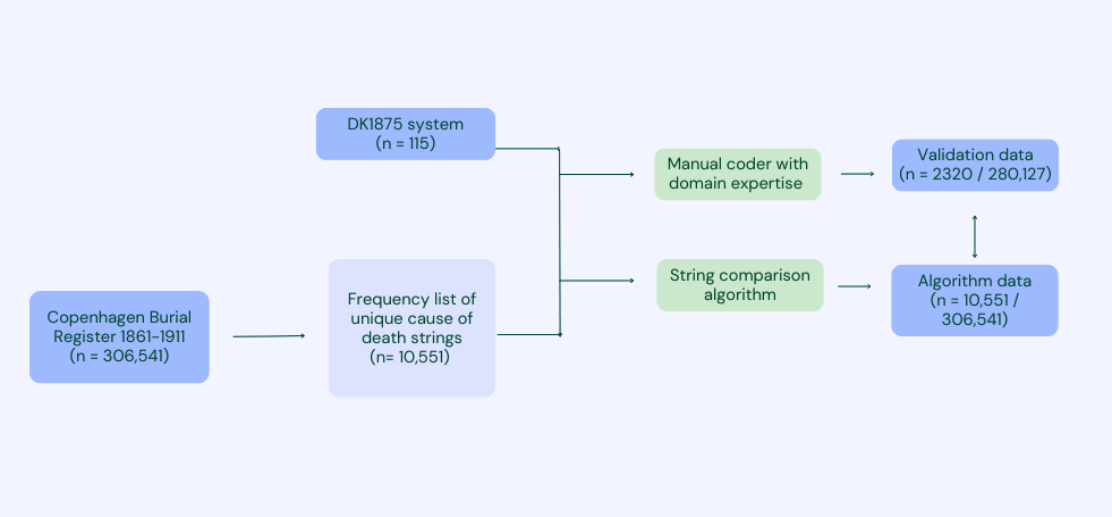

Figure 1: Diagram of the sources used in this paper.

focuses heavily on automated coding to the modern ICD-10 system, which does not give us any information on how an automated coding algorithm would work within the context of a historical coding and classification system.

This paper aims to contribute to the discussion on automated coding by exploring how accurately we can code historical causes of death to a historical classification system with the relatively simple technique of string similarity matching. Using a historical classification system is essential for string similarity techniques, as a modern system such as the ICD10 of ICD10h [14] will either be in a different language or have very different phrasings compared to the contemporary causes of death on the death certificates. We therefore apply a set of automated string similarity algorithms to code our cause of death data to "DK1875", a coding and classification system developed for use by nineteenth century physicians in Denmark (see appendix 1 for the full DK1875 system). This is done without any previous training of the algorithm, and thus without any use of training data. Since the causes of death in the Copenhagen Burial Register have already been manually coded to DK1875, we compare the performance of the algorithm to this validation dataset, constructed by a manual (historian) coder with extensive domain expertise (Figure 1).

The paper is based on work done in a 48-hour hackathon at UiT The Arctic University of Norway in May 2022, in which all authors participated. At the hackathon, two small teams were competing to solve the following challenge: code the 10,000+ unique causes of death in the Copenhagen Burial Register to the correct category in the DK1875 system. In the hackathon and in this paper, we have defined 'correct' coding as one that matches the coding done by hand in the validation data. To tackle the issue, both teams were given 1) a frequency list of all unique causes of death strings in the burial data and 2) a list of the 115 categories of the DK1875 system. Both teams chose string similarity matching to assign a code to the causes of death, but their methods differed slightly. One team did only little data cleaning, relying

instead on a range of different string similarity methods to find the best fit for each cause of death based on confidence scores. The second team did more extensive data cleaning and relied upon a single-string similarity measure. In this paper, we rely on the latter approach to improve automatic coding and study which possibilities and limitations it creates, as it seems to provide the most consistent results[3].

2. The data

2.1. The Copenhagen Burial Register

In 1861 the Copenhagen municipality introduced the Copenhagen Burial Register to centralize the administration of the growing number of burials in the city caused by rapid growth in population. Previously the burials had been recorded in individual registers for each cemetery, but with the introduction of the Copenhagen Burial Register, all burials within the city (except for the burials at military, Roman-Catholic and Jewish cemeteries) were to be recorded in one volume. From 1887 the registration rules were changed, and from then on, the Copenhagen Burial Register contains all deaths that happened in the city, regardless of where the burial took place [20]. The Burial Register is remarkably consistent; no volumes are missing, and they contain close to the same information for each burial throughout the period. The information includes name, age, occupation, date of death, cause of death, burial date and place, address at death and where the deceased's body was kept prior to burial [21]. Earlier studies have compared the information to that found in the equivalent death certificates and concluded that while the information is mostly the same, the Copenhagen Burial Register provides more complete personal information [22].

The individual-level burial records contained in the register have first been digitized, and then transcribed by volunteers at the Copenhagen City Archives, resulting in scanned facsimiles and a machine-readable version with a link between the two versions of the same burial [20]. For the period from 1861 to 1911, the database contains a total of 306,541 burials and 10,551 unique causes of death. When transcribing the causes of death, volunteers were instructed to do it "…i den rækkefølge de står og skrives kildetro, på nær forkortelser, variationer i stavemåder og åbenlyse stavefejl." ["…in the order they appear and true to the source, except for abbreviations, variations in spelling and obvious mistakes in spelling"] [23]. The transcriber had the option to either choose from a drop-down list of previously used cause-of-death values in the source or to write a free-text alternative themselves if the drop-down list did not contain the cause of death they were transcribing. Finally, a quality control procedure handled by so-called 'super-users' has corrected errors and ensured that the transcription always corresponds to the handwritten word or string on the record. All these initiatives contributed to standardizing and cleaning the causes of death as a part of the transcription process while interfering as little as possible with the content of the source. As a result, the transcribed causes of death from the Copenhagen Burial Register include different spellings for the same word into one string, such as "ukendt (ubekendt, ubekjendt)" ["unknown"], while separating synonyms for

[3]Python scripts and data used for the analysis in this paper are available in the following GitHub repository link: https://github.com/louiseludvigsen/Automated-Coding-of-Historical-Danish-Cause-of-Death-Data.

Table 1
Example of the frequency list of unique cause of death strings

ID	Tidy cod	Freq	Acc.sum	Acc.perc
1	morbus cordis (mb. cordis, mb. cord.)	15,100	15,100	3.44
2	dødfødt	13,290	28,390	6.47
3	pneumonia (pneumoni)	11,813	40,203	9.16
4	..	..	..	..

the same illness, for example: “Engelsk syge” [Danish term for rachitis] and “Rachitis”. Further, if a particular cause of death has a descriptive attachment, it is listed as a separate string, for example: “Nephritis” and “Nephritis chronica”. Overall, the transcription setup of the archives has lessened the need for data cleaning, but some issues remain for doing a string comparison, as will be explained in the methods section.

2.2. The frequency list of unique cause of death strings

A frequency list of all the unique cause-of-death strings that were used for the 306,541 burials in the Copenhagen Burial Register 1861-1911 was created by Ludvigsen, without considering whether the cause-of-death string was used as first, second, third etc. cause of death. The list contains 10,551 rows (one for each unique cause of death string) with the following information:

- tidy cod: the unique cause of death string
- Freq: frequency of use as the primary cause of death
- Perc: The percentage of burials with this string as the primary cause of death
- Acc. Freq: accumulated frequency of use as the primary cause of death
- Acc. Perc: accumulated percentage of burials with these strings as the primary cause of death.

This list was used unconnectedly for the automated string comparison and for Ludvigsen to create the validation data.

2.3. The validation data

The validation data was created based on the frequency list described above, by Louise Ludvigsen, a historian with large domain expertise in the registration and coding of historical causes of death in nineteenth- and twentieth-century Denmark. Ludvigsen manually coded 2,320 of the 10,551 unique cause of death strings in the list, by order of frequency starting with the most frequent, using about 6 months to do so. The 2,320 manually coded cause of death strings are used as the primary cause of death for 280,127 of the 306,541 burials (91.38%).

Each of the 2,320 unique cause of death strings was coded to three different coding systems in two separate steps. First, the unique cause of death was coded to the ICD10h coding scheme, constructed by the SHiP network, adapting an earlier coding scheme developed by the Cambridge Group for the History of Population and Social Structure [14]. The ICD10h is largely based on

the ICD10 system designed by the World Health Organization (version 2016), but it contains two additional digits at the end of each category to allow for more historical nuance (e.g. A16.904) [14, p. 68].By default, the causes were assigned ICD10 codes as well since they are essentially identical to ICD10h save the suffix. Secondly, the same unique cause of death string was coded to the contemporary Danish system DK1875, used in the official Danish cause of death statistics from 1876-1930 [24, p.188]. As mentioned previously, we only make use of the DK1875 classification system in this paper. Using a modern system such as the ICD10 of ICD10h will not work for our use of a string similarity technique, as they will either be in a different language or have very different phrasings compared to the contemporary causes of death on the death certificates.

2.4. The DK1875 system

The DK1875 system was introduced in the official Danish cause of death statistics on the 1st of January 1875 and consists of 115 categories, placed within nine main groups (see appendix A). The 115 categories describe either a singular disease or groups of diseases, such as “Kighoste” [“Whooping cough”] and “Andre Farsoter” [“Other epidemic diseases”]. Each category has a number, a Latin description, and a Danish description (see appendix A). The system was inspired by the contemporary English and Swedish systems, and particularly by the principles of the English statistician Farr [24, p.165]. It is multifaceted as it works simultaneously as a nomenclature, a coding scheme, and a classification system. It serves as a nomenclature, as the physicians were instructed to use the 115 categories when filling out the death certificates [25], and as a coding and classification scheme when used in the official cause of death statistics and other statistical analysis, either as the 115 categories or added together into larger groups for analysis. The DK1875 system received only very minor updates before it was ultimately replaced by a common Nordic system in 1931 [24, p. 193].

3. Methods

Based on our experiences from the hackathon at UiT The Arctic University of Norway, we decided to apply a single method for the automated coding in this paper. We have employed the Jaro-Winkler string similarity measure, which has been shown to be effective in matching strings with typographical errors and inconsistencies. The Jaro-Winkler measure is based on the Jaro distance, which calculates the similarity between two strings based on the number of matching characters and the number of transpositions required to make the strings identical. The Winkler modification adds a scaling factor based on the length of the common prefix of the strings, which accounts for the likelihood that two similar strings have a common prefix. Due to these features, we expect Jaro-Winkler to perform better than other string similarity measures, such as Levenshtein distance and cosine similarity which only captures edit distance, since Jaro-Winkler is more robust to differences in string length and minor typographical errors.

3.1. Data cleaning and string comparison

There were three parts to the data cleaning in the frequency list and the DK1875 list: abbreviations, information in parentheses and translating the Nordic letters æ, ø and å. In the frequency list, we removed the accents and special characters from the unique cause of death strings, as well as transliterating the Nordic letters æ, ø and å to ae, o and a, respectively. In addition, we moved the contents of the parentheses in the unique cause of death strings to a new column since they would otherwise disturb the string-matching algorithm. The information in the parentheses either contained an abbreviation, such as "tuberculosis pulmonum (tub. pulm.).", or the Latin phrase for a cause written in Danish (or vice versa). In the DK1875 coding scheme, cleaning was done for both the Latin and Danish terms for each code. First, abbreviations were written out, e.g. "bronchit." became "bronchitis". Secondly, we shortened categories with lengthy names when we suspected it might improve the matching. And finally, we separated categories when they contained several specific causes of death, such as "Cholerine & catarrhus intestinales," ["Domestic cholera and acute diarrhoea"], which were separated into "Cholerine" and "catarrhus intestinalis".

After cleaning the data, we used the 'pyjarowinkler' package [26] in Python to calculate the Jaro-Winkler distances, which returns a score between 0 and 1, with 1 being the highest similarity between the two strings. For each unique cause of death string, we computed Jaro-Winkler distances between the string and both the Danish and Latin labels of each of the 115 DK1875 categories. The string was then assigned the DK1875 code with the highest density score. No minimum score was set for assigning a code, and thus all 10,551 unique cause of death strings were assigned a code.

3.2. Testing the accuracy

To test the accuracy of our approach and the consequences it may have for studies on historical mortality, we have compared the codes assigned by the automated string comparison method to the validation data in three different ways: at the level of each unique cause-of-death string, at the level of each Heiberg group and for the overall cause-of-death pattern for all burials in Copenhagen 1861-1911.

To test the accuracy of our approach at the level of each unique cause of death string, we compared the DK1875 code assigned by the string comparison algorithm directly to the manually assigned DK1875 codes in the validation dataset which is our "ground truth" data. If the two codes were the same, we considered the code assigned by the algorithm correct. If they were different from each other, we considered the code assigned by the algorithm wrong. As not all 10,551 strings have been coded by hand, the comparison was based on the 2320 strings that were.

However, in practice, the 115 subgroups are rarely used directly for analysis. Since there are too many of them, and some describe groups of diseases while others describe individual diseases, it is more convenient to classify them into larger groups. In this paper, we have used the classification system presented by the Danish physician Povl Heiberg in his study of mortality among 15–74-year-olds in Denmark in the 1890s and early 1900s (from here on referred to as the Heiberg groups)[27]. Heiberg himself assigned each of the 115 DK1875 subgroups to one of

twelve groups in his classification scheme (see appendix B). We are interested in two measures of coding accuracy here: 1) the proportion of the codes in each Heiberg group that have been coded accurately according to the validation data. This is to see if certain disease groups are more difficult to code for the automatic method. 2) the proportion of the individual-level causes of death that have been assigned to the correct Heiberg group, according to the validation data, even though the DK1875 might be inaccurate. This will help us measure when the algorithm has not coded to the same code as the validation data, but to an adjacent category that may be very similar. Using the Heiberg groups allows us to look closer at what effects the accuracy will have for studies on historical mortality, which often makes use of larger groups like these for analysis.

Finally, we have examined how the automated coding algorithm performs when analyzing the cause of death patterns for the burials of people aged 0-80 in Copenhagen 1861-1911. For this analysis only, the burials of people over age 80 have been excluded, and the Heiberg groups "suicides", "accidents" and "genitourinary diseases" were put into the Heiberg group "others".

4. Results

4.1. Coding of the unique cause of death strings

Of the 2,320 coded causes of death in the validation data, our algorithm managed to code 1,075 (46.3%) strings correctly (i.e. in the same way as the manual coder). If we look at the number of burials coded correctly, rather than the unique causes of death, the number is somewhat higher, as the algorithm was able to code 176,681 of the 280,127 burials (63%) correctly compared to the validation data. There are more correct codes for the burials than there are at the level of each unique cause of death since the automated approach overall tends to be slightly more accurate in coding the causes that most frequently appear in the burials. Ordered by frequency, 283 of the 500 most frequent causes were coded accurately (56%), while the number was 667 out of the remaining 1,529 causes of death (44%). This probably reflects the fact that the more frequent causes tend to be shorter and simpler, while the less frequent causes are often more elaborate in detail, and thus differ more from the categories in string similarity.

The 100 most common unique causes of death strings (roughly 1% of all the total unique cause of death strings) account for more than 78% of all the burials in Copenhagen from 1861-1911. Among these 100, the automated string comparison was able to assign a correct code for 65 out of 100 of the unique cause of death strings compared to the validation data (Figure 2). The accurately coded cause of death strings tends to be relatively short and have a very high overlap with the wording in the DK1875 descriptions. They are also mostly distinctive causes of death in the sense that there is only one category in the DK1875 system that will match, and not several for the algorithm to choose between. Smallpox, for instance, has a single code that all cases are assigned to. Tuberculosis, on the other hand, has several codes in the DK1875 system, accounting for different variations of the disease: "29: Akut miliærtuberkulose" [acute miliary tuberculosis], "30: lungevindsot" [pulmonary phthisis] and "31: tuberkulose i andre organer" [tuberculosis in other organs]. For a simple algorithm like the string similarity method applied here, it is very difficult to distinguish between these.

For the top 10 most frequent unique cause-of-death strings, the string comparison algorithm

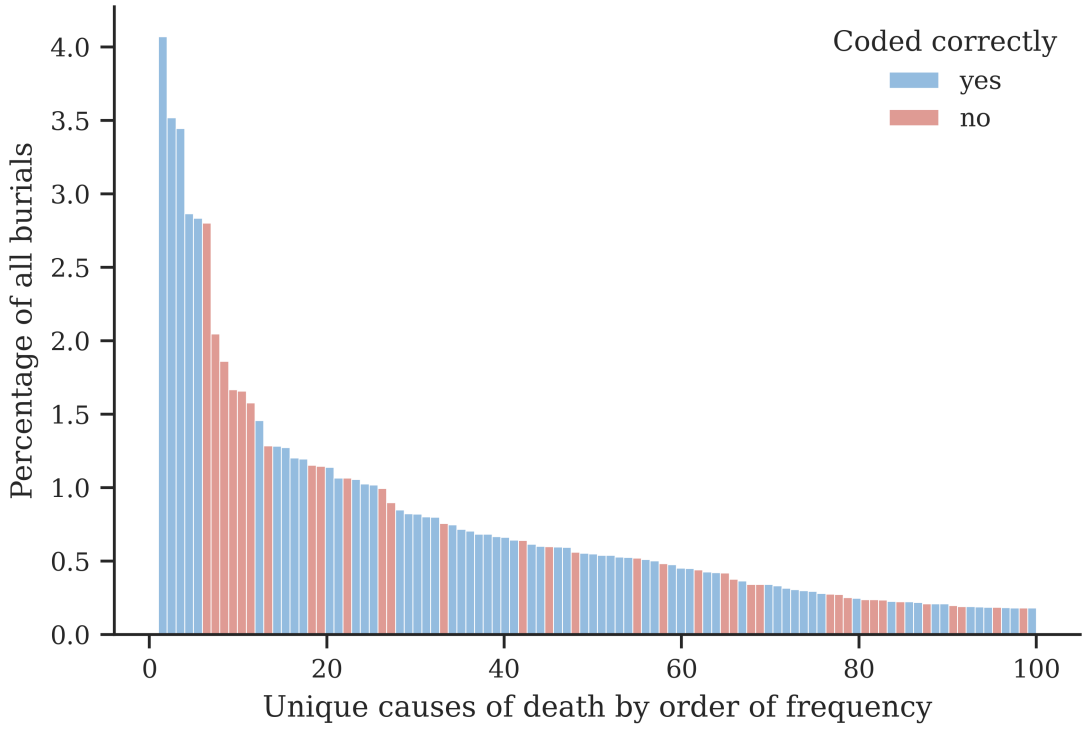

Figure 2: Automated string comparison coding compared to validation dataset, for the 100 most frequent causes of death in the Copenhagen Burials 1861-1911.

assigns a wrong code in half of the cases (Figure 2). Most of these are relatively easy to code manually, but the algorithm gets them wrong because the strings are completely different in the frequency list and the DK1875 system. This is the case for the strings “kramper” ["cramps"], “ukendt” ["unknown"], and “pludselig død” ["sudden death"] in the frequency list. “Kramper” [“cramps”] is a Danish term for “convulsioner” ["convulsions"], which has an exact match in the DK1875 categories. Similarly, the proper code for “ukendt” ["unknown”] would be no. 113, “Uangivet eller slet specificeret dødsårsag” ["Unnoted or poorly specified cause of death"]. In this case, the language of the DK1875 system is not only too different, but also too elaborate for a string similarity method to work. In other cases, the automated method fails because the causes in question fit into several categories, once again caused by the DK1875 system being too elaborate. This is the case for “tuberculosis pulmonum” [“pulmonary tuberculosis”], “meningitis” and “bronchitis”. Tuberculosis has three different categories in the DK1875 system as mentioned previously, while meningitis has two different categories, “61, hjernebetændelse” [“cerebral meningitis”] and “69, Rygmarvsbetændelse” [“spinal meningitis”], which both contain the word meningitis in the Latin phrasing. Bronchitis has three different categories, which all contain the word bronchitis, but distinguish between acute, chronic, and capillary bronchitis: “76: Brystkatarrh, akut bronchitis” ["Chest catarrh, acute bronchitis"], “77: Kapillær Bronchitis og katarrhalsk Lungebetændelse” ["Capillary Bronchitis and catharrhal pneumonia"], “78, Chronisk

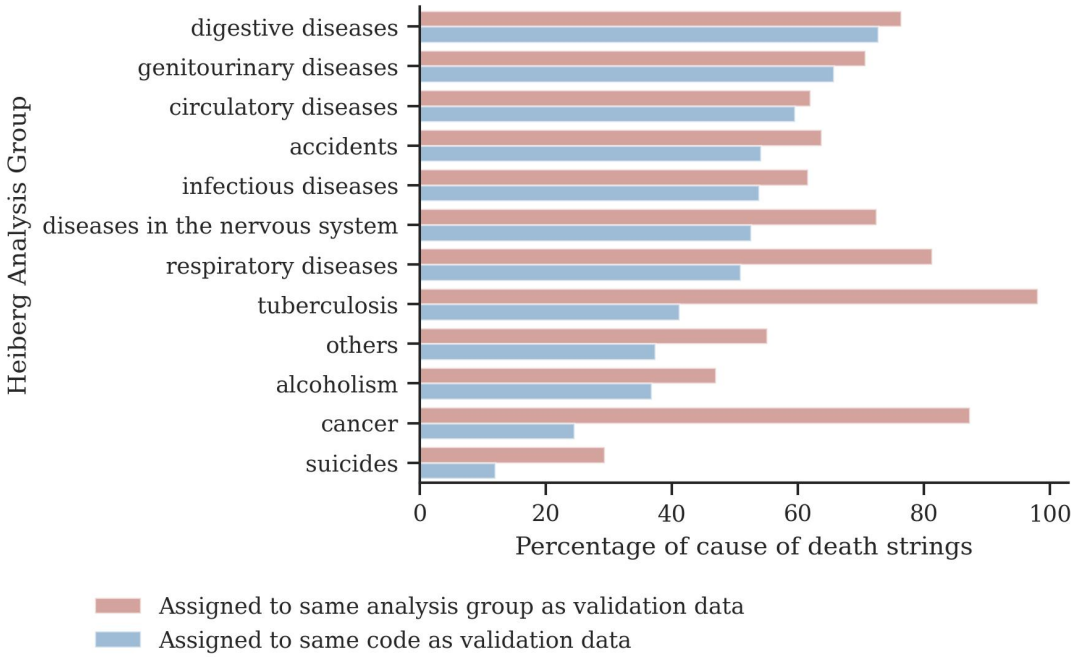

Figure 3: Measures of accuracy for each group in the Heiberg scheme.

Bronchitis" ["Chronic bronchitis"]. All three examples confirm that it is very difficult for a simple string similarity method to distinguish between multiple categories for variations of the same disease.

4.2. Coding according to Heiberg groups

As mentioned in the methods section, we are not only interested in how the string comparison method performs for each unique cause of death string, but also how it performs when looking at larger groups for analysis, such as the Heiberg groups. To see if certain disease groups are more difficult to code for the automatic method, we look at the proportion of the codes in each Heiberg group that have been coded accurately according to the validation data. For this code-specific accuracy, that is the proportion of the cause of death strings that the automated method has coded to the same code as the validation data, there is quite a bit of variation between the disease groups (Figure 3). While the automated method performs relatively well for digestive, circulatory, and genitourinary diseases, with an accuracy of c. 60-70%, it performs poorly for suicides, tuberculosis, and cancer. It is not surprising that suicides are difficult for the algorithm to code since the burial records most often contain a description of the manner of the suicide, rather than the word itself.

The pattern is different if we look at accuracy according to Heiberg's groups, meaning causes

of death where the algorithm has hit the same Heiberg group as the validation data, regardless of the specific DK1875 code. For tuberculosis, the accuracy is almost 100% at the group level, compared to 'just' 40% at the code level. Similarly, cancer diseases, for which the algorithm performs very poorly at the code level, are coded to the right Heiberg group almost 90% of the time (Figure 3). This striking difference in accuracy between the DK1875 codes and the Heiberg groups reflects that while the codes are wrong according to the validation data, they are coded to an adjacent, and probably very similar, code, and thus end up in the same classification group. For instance, the most common code-level error for cancer diseases is when the algorithm has placed a cause of death in codes like "Breast cancer", "Stomach cancer" or "Cervical cancer" when the appropriate is a more general one: "Cancer in other organs". This suggests that overall, the automated method is actually very accurate at capturing causes of death related to cancer, but due to the detail of the coding scheme, it is often mistaken in the individual code. The reason why the code- and group-specific accuracy for tuberculosis varies so much is the same: since there are a handful of tuberculosis-related codes in the DK1875 system, the algorithm often picks the wrong one, because it chooses the one with the shortest edit distance, but it rarely picks one that is unrelated to tuberculosis.

For the categories where the code- and group-level accuracy is similar, like digestive and circulatory diseases, this reflects that when the algorithm has chosen the wrong code, it is usually very far off, and not just another code within the same Heiberg group. For example, within the Heiberg group of digestive diseases several of the unique cause-of-death strings that in the validation data have been coded to intestinal ruptures, have been coded to both cancers and bronchitis with the string comparison method. Likewise, the most common faulty coding of heart disease is to encephalitis.

4.3. Coding according to the cause of death patterns

Overall, the cause of death pattern based on the codes assigned by the string comparison method looks fairly similar to the pattern based on the manual codes. This is very remarkable, considering that the string comparison method only assigned a correct code for 63% of the burials compared to the validation data (Figure 4).

However, on closer inspection, there are several clear differences between the patterns produced by the two methods. The cause of death pattern based on the string comparison method has more burials coded to alcoholism and cancer than the manual method. This is probably because the cause of death strings for alcoholism and cancer is often written in a multitude of ways (all containing the word alcoholism or cancer), making most of them appear quite far down the frequency list, and thus less likely to be coded by the manual coder. In addition, the automated method also has more burials coded to respiratory diseases, digestive diseases, infectious diseases, and diseases in the nervous system, but fewer burials coded to other diseases. These differences are particularly clear in the first and last years of the period. This might be because the classification system used is from 1876, which means the years before are likely to have slightly different phrasings for the causes of death, which could also explain the final years, where new knowledge and diagnoses might be hard to match with the old classification system. In combination, these differences mean that when we analyse the cause of death patterns and their development over time, the results are quite different for the

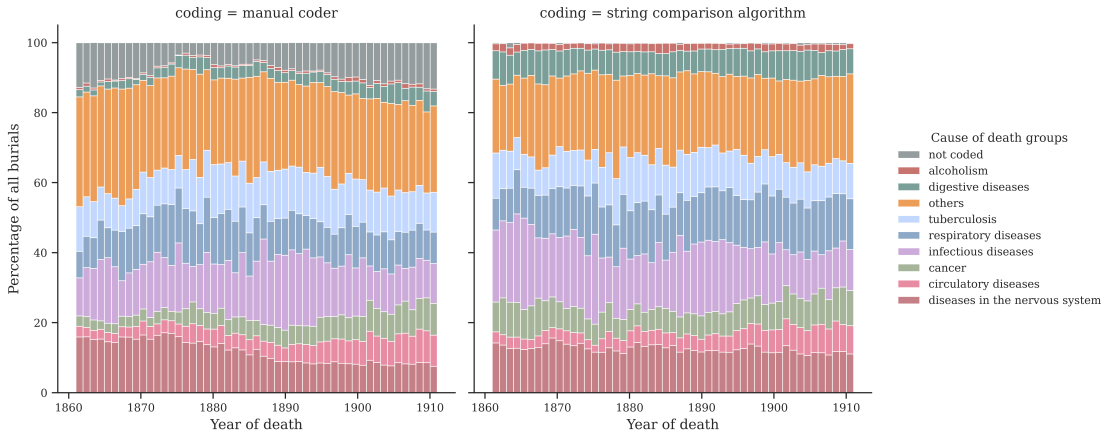

Figure 4: Cause of death pattern for the ages 0-80 in Copenhagen 1861-1911.

two methods, with the pattern from the string comparison method seeming much more stable throughout the period.

Some of the differences between the two cause of death patterns may be explained by the fact that the string comparison algorithm has coded all burials, whereas the manual coder has only coded 91.38% of the burials. However, the results from the analysis of the Heiberg groups and the unique causes of death has shown that only 46.3% of the unique causes of death are coded correctly, and that some causal groups are more affected than others, which suggest that the differences are more likely to be because of wrongly assigned codes.

5. Concluding remarks

The aim of this paper was to explore how accurately we can code historical causes of death to a historical classification system with the relatively simple technique of string similarity matching. We found that our algorithm managed to code 46.3% of the unique cause of death strings and 176,681 (63%) of the burials correctly compared to the validation data. Examining the coding of each unique cause of death string, the automated method performs relatively well for digestive, circulatory, and genitourinary diseases, with an accuracy of c. 60-70%, it performs poorly for suicides, tuberculosis, and cancer. However, for tuberculosis and cancer, the algorithm is very good at assigning a code within the same Heiberg group as the validation data, even if the individual code for the unique cause of death is wrong. For tuberculosis, the accuracy is almost 100% at the group level, compared to 'just' 40% at the code level. Similarly, cancer diseases, for which the algorithm performs very poorly at the code level, are coded to the right Heiberg group almost 90% of the time. Finally, we find that the cause of death pattern based on the codes assigned by the string comparison method is fairly similar to the pattern based on the manual codes, which is impressive since the string comparison method only assigns a correct code to 46.3% of the unique cause of death strings. However, upon closer inspection there are several discrepancies between the two patterns which would result in two quite different results when

analysing the cause of death patterns and their development over time. It seems most likely that these discrepancies have occurred because of the high percentage of wrongly assigned codes amongst the unique causes of death, and the fact that certain Heiberg groups are more affected by this than others.

From examining the automatic coding of the unique cause of death strings, we can point out two major flaws of the string similarity method. Firstly, a number of the relatively straightforward causes of death at the top of the frequency list were coded inaccurately according to the validation data simply due to significant differences in the wording between the two strings. Examples include "Kramper" ["cramps"], "ukendt" ["unknown"] and "pludselig død" ["sudden death"]. These issues could be addressed by adding synonyms to the coding scheme, so that it has more than one label for each category, and thus would be more flexible to the different but semantically identical terms. Secondly, in other cases, the automated method fails because the unique cause of death string in question fits into several categories in the DK1875 system. It is difficult to increase the accuracy of the string comparison method in these cases, since they require contextual knowledge. Even a manual coder will often need additional help such as clinical dictionaries and medical literature to know more about the range of distinct expressions of the same disease.

When we consider these issues, the version of the method applied here is not accurate enough to use for actual data analysis of the cause of death patterns. While it is encouraging that the string similarity comparison is close to presenting a somewhat accurate representation of the overall distribution cause of death groups, we could paint the same picture using aggregate cause-of-death statistics, which defeats the purpose of working on individual-level data. We do find that while some of the causes that were inaccurately coded were in fact coded to another category within the same Heiberg group, meaning that it was coded to an adjacent illness, others were coded to completely different categories. This would present a major issue if the data was used to study the development of specific causes. We would argue that scholars are increasingly interested in doing just that: analyses of specific illnesses, in connection to or rather than the overall panorama of disease.

The promise of automated coding is that it can help cover the diversity of causes, by coding the long tail of cause of death strings that only appear once or twice in the dataset, which a manual coder will rarely reach. Based on the method's performance on the validation data, we would not trust it to work well on a large dataset with no validation checks. However, it should be taken into account that the algorithm used was developed quickly during a 48-hour hackathon. Our results represent a developer effort that is realistic for individual research projects, and our code could be adjusted for other projects. The results of the paper provide a baseline, and it is very well possible to increase its accuracy. A way forward could be to work with the cut-off points of the Jaro-Winkler scores, setting a threshold so that only those causes where the string similarity match is relatively certain are coded and leaving the remainder to a manual coder. While doing so would ensure a higher degree of certainty in the code assigned, it would reduce the number of causes of death coded. In addition, the amount of uncoded causes of death would most likely not be equally distributed amongst all cause of death groups, meaning that even though the certainty of each code is higher, the sample in total may be more biased. As it is now, the automated coding would perhaps be better used as a method to catch most of the initial cases of a certain disease with a very set phrasing, such as cancer. Afterwards, the

researcher could look for more cases, such as tumours. In this way, it would be a helpful tool for locating cases of the specific disease for a study of this disease. However, this only works for certain types of diseases, where the word is consistently used in both the coding and the unique cause of death string.

Acknowledgments

Many thanks to the Copenhagen City Archives and their volunteers for the creation of the Copenhagen Burial Register Dataset, without which none of this would have been possible.

References

[1] A. R. Omran, The epidemiologic transition: a theory of the epidemiology of population change., The Milbank Quarterly 83 (2005) 731–757. doi:10.1111/j.1468-0009.2005.00398.x.

[2] T. McKeown, The modern rise of population, Academic Press, Nueva York, 1976.

[3] A. Løkke, Døden i barndommen: spædbørnsdødelighed og moderniseringsprocesser i Danmark 1800 til 1920, Gyldendal, København, 1998.

[4] A. Reid, E. Garrett, C. Dibben, L. Williamson, '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) 320–344. URL: https://doi.org/10.1080/1081602X.2014.1001768. doi:10.1080/1081602X.2014.1001768, number: 3.

[5] J. P. Mackenbach, The Epidemiologic Transition Theory, Journal of Epidemiology and Community Health (1979-) 48 (1994) 329–331. URL: http://www.jstor.org/stable/25567930, publisher: BMJ.

[6] G. C. Alter, A. G. Carmichael, Classifying the Dead: Toward a History of the Registration of Causes of Death, Journal of the History of Medicine and Allied Sciences 54 (1999) 114–132. URL: http://www.jstor.org/stable/24624555.

[7] Det Statistiske Bureau, Statistisk Tabelværk, Fjerde række, Litra A Nr. 2, Vielser, Fødsler og Dødsfald i Aarene 1875-1879 samt Dødsaarsagerne i aarene 1876-1879., Bianco Lunos Bogtrykkeri, Kjøbenhavn, 1882.

[8] G. B. Risse, Cause of death as a historical problem, Continuity and Change 12 (1997) 175–188. URL: http://www.cambridge.org/core/journals/continuity-and-change/article/cause-of-death-as-a-historical-problem/74166C12EDE0C2B9C9AD025DD85948E4. doi:10.1017/S0268416097002890, number: 2.

[9] G. C. Alter, A. G. Carmichael, Reflections on the classification of causes of death, Continuity and change 12 (1997) 169–173. doi:10.1017/S0268416097002889, place: Cambridge Publisher: Cambridge University Press.

[10] A. Janssens, I. Devos, The Limits and Possibilities of Cause of Death Categorisation for Understanding Late Nineteenth Century Mortality, Social History of Medicine 35 (2022) 1053–1063. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9949560/. doi:10.1093/shm/hkac040.

[11] J. Arrizabalaga, Medical Causes of Death in Preindustrial Europe: Some Historiographical

Considerations, Journal of the History of Medicine and Allied Sciences 54 (1999) 241–260. URL: http://www.jstor.org/stable/24624562.

[12] S. J. Kunitz, Premises, Premises: Comments on the Comparability of Classifications, Journal of the History of Medicine and Allied Sciences 54 (1999) 226–240. URL: http://www.jstor.org/stable/24624561.

[13] G. C. Alter, A. G. Carmichael, Studying causes of death in the past : problems and models, Historical methods 29 (1996) 44–48.

[14] A. Janssens, Constructing SHiP and an International Historical Coding System for Causes of Death, Historical Life Course Studies 10 (2021) 64–70. URL: https://hlcs.nl/article/view/9569. doi:10.51964/hlcs9569.

[15] Københavns begravelsesprotokoller 1861–1911 (n=306.541) [The Copenhagen burial register 1861–1911], October 2020. URL: https://www.rigsarkivet.dk/udforsk/link-lives-data/, type: dataset.

[16] P. Harteloh, The implementation of an automated coding system for cause-of-death statistics, Informatics for health & social care 45 (2020) 1–14. doi:10.1080/17538157.2018.1496092, place: England Publisher: Taylor & Francis.

[17] J. Carson, G. Kirby, A. Dearle, L. Williamson, A. Reid, C. Dibben, "Exploiting historical registers: Automatic methods for coding 19th and 20th-century cause of death descriptions to standard classifications"., in: New Techniques and Technologies for Statistics, Eurostat, 2013, pp. 598–607.

[18] B. Koopman, G. Zuccon, A. Nguyen, A. Bergheim, N. Grayson, Automatic ICD-10 classification of cancers from free-text death certificates, International Journal of Medical Informatics 84 (2015) 956–965. URL: https://www.sciencedirect.com/science/article/pii/S1386505615300289. doi:10.1016/j.ijmedinf.2015.08.004.

[19] K. Xu, M. Lam, J. Pang, X. Gao, C. Band, P. Mathur, F. Papay, A. K. Khanna, J. B. Cywinski, K. Maheshwari, P. Xie, E. P. Xing, Multimodal Machine Learning for Automated ICD Coding, in: Proceedings of the 4th Machine Learning for Healthcare Conference, PMLR, 2019, pp. 197–215. URL: https://proceedings.mlr.press/v106/xu19a.html, iSSN: 2640-3498.

[20] K. Stadsarkiv, Begravelser i hele København 1861-1942, 2022-05-02. URL: https://kbharkiv.dk/brug-samlingerne/kilder-paa-nettet/begravelser-i-koebenhavn/begravelser-1861-og-frem/.

[21] L. Ludvigsen, B. Revuelta-Eugercios, A. Løkke, Cause-Specific Infant Mortality in Copenhagen 1861–1911 Explored Using Individual-Level Data, Historical Life Course Studies 13 (2023) 9–43. URL: https://hlcs.nl/article/view/12032. doi:10.51964/hlcs12032.

[22] B. Revuelta-Eugercios, H. Castenbrandt, A. Løkke, 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 : the journal of the Society for the Social History of Medicine (2021). doi:10.1093/shm/hkab037.

[23] K. Stadsarkiv, Indtastningsvejledning - Begravelser 1861-1940, 2023-03-17. URL: https://kbharkiv.dk/deltag/indtast-begravelser-1861-1940/indtastningsvejledning-begravelser-1861-1940/.

[24] B. Johansson, Den danske sygdoms- og dødsaarsagsstatistik : Med et afsnit om pneumonistatistik., Ejnar Munksgaard, Kbh, 1946.

[25] Anvisning for Læger med Hensyn til Udstedelsen af Dødsattester., Det Kongelige Sundheds-

Collegium, 1875. URL: http://www5.kb.dk/e-mat/dod/130021307433.pdf.
[26] J.-B. Ratte, pyjarowinkler 1.8, March 23, 2016. URL: https://pypi.org/project/pyjarowinkler/.
[27] P. Heiberg, Dødeligheden og Dødsaarsagerne i Danmark i de 2 Tiaar 1890-1899 og 1900-1909 i Aldersklasserne 15-74 Aar., Særtryk af Bibliotek for Læger, 1918.

A. Appendix A: The DK1875 system

Nr.	Latin	Danish
I	Morbi epidemici	Farsoter
1	Variolæ	Kopper
2	Morbilli	Mæslinger
3	Scarlatina	Skarlagensfeber
4	Diphtheritis	Ondartet Halssyge
5	Croup	Strubehoste
6	Tussis convulsiva	Kighoste
7	Febris tyhoidea	Typhoid Feber
8	Typhus exanthematicus	Exanthematisk Typhus
9	Dysenteria	Blodgang
10	Cholera asiatica	Asiatisk Cholera
11	Cholerine & Catarrhus intest. acutus	Indenlandsk Cholera og akut Diarrhoe
12	Erysipelas faciei & ambulans	Ansigts- og anden Vandrerosen
13	Febris puerperalis	Barselfeber
14	Pyæmia & Septichæmia	Ondartet Saarfeber
15	Febris intermittens	Koldfeber
16	Influenza	Grippe
17	Febris rheumatica	Akut Ledderheumatisme
18	Alii morbi epidemici	Andre Farsoter
II	Sangvinis infectiones	Blodforgiftninger
19	Malleus humidus	Snive
20	Pustula maglina	Miltbrand
21	Alia venena animalia	Andre dyriske Gifte
22	Syphilis acquisita	Erhvervet Syphilis
23	Syphilis congenita	Medfødt Syphilis
24	Alcoholismus chronicus	Brændeviinssygdom
25	Delirium tremens	Drankergalskab
26	Mors in ebrietate	Pludselig Død af Drik
III	Morbi Constitutionales	Konstitutionelle sygdomme
27	Scrophulosis	Kirtelsyge
28	Hydrocephalus acutus	Akut Hjernevandsot
29	Tuberculosis acuta	Akut Miliærtuberkulose
30	Phthisis pulmonum	Lungesvindsot

31	Tuberculosis in aliis corporis partibus	Tuberkulose i andre Organer
32	Cancer ventriculi	Mavekræft
33	Cancer uteri	Livmoderkræft
34	Cancer mammæ	Brystkræft
35	Cancer in aliis corporis partibus	Kræft i andre Organer
36	Rhachitis	Engelsk Syge
37	Diabetes mellitus	Sukkersyge
38	Scorbutus	Skjørbug
39	Anæmia	Anæmi
IV	Violentæ mortis causæ	Voldsomme dødsårsager
40	Casus Mortiferi: Præcipitatio & Contusio	Ulykkelige Hændelser: Fald og Knusning
41	Casus Mortiferi: Submersio	Ulykkelige Hændelser: Drukning
42	Casus Mortiferi: Suffocatio	Ulykkelige Hændelser: Kvælning og Ihjelliggen
43	Casus Mortiferi: Vulnus sclopetarium	Ulykkelige Hændelser: Skudsaar
44	Casus Mortiferi: Vulnus incisum & punctum	Ulykkelige Hændelser: Snit- og Stiksaar
45	Casus Mortiferi: Ambustio	Ulykkelige Hændelser: Forbrænding og Skoldning
46	Casus Mortiferi: Congelatio	Ulykkelige Hændelser: Forfrysning
47	Casus Mortiferi: Veneficium	Ulykkelige Hændelser: Forgiftning
48	Casus Mortiferi: Alii casus mortiferi	Ulykkelige Hændelser: Andre ulykkelige hændelser
49	Suicidium: Submersio	Selvmord: Drukning
50	Suicidium: Strangulatio	Selvmord: Hængning
51	Suicidium: Vulnus sclopetarium	Selvmord: Skud
52	Suicidium: Vulnus incisum & punctum	Selvmord: Snit og stik
53	Suicidium: Venenum	Selvmord: Gift
54	Suicidium: Alii suicidii modi	Selvmord: Andre Selvmords Dødsaarssager
55	Homicidum	Mord og Drab
V	Vitia innata	Dannelsesfejl
56	Cyanosis	Blaasot
57	Debilitas congenita	Medfødt Svaghed
58	Spina bifida	Medfødt Rygmarvsvandsot
59	Atelectasis	Mangelfuld Udvidning af Lungerne
60	Alia vitia innata	Andre Dannelsesfejl
VI	Morbi singulorum organorum	Lokale organsygdomme
61	Encephalitis et Meningitis cerebralis	Hjernebetændelse
62	Apoplexia cerebri	Apoplexi
63	Morbi cerebri chronici	Chroniske Hjernesygdomme
64	Morbus mentalis	Sindssygdom
65	Tetanus	Stivkrampe
66	Trismus	Mundklemme
67	Epilepsia	Ligfald
68	Ecclampsia	Konvulsioner

69	Myelitis & Meningitis spinalis	Rygmarvsbetændelse
70	Ataxia s. Tubes dorsalis	Rygmarvstæring
71	Alii medullæ spinalis morbi chronici	Andre chroniske Rygmarvssygdomme
72	Laryngitis	Strubebetændelse
73	Morbi laryngis chronici	Chroniske Strubesygdomme
74	Pneumonia	Lungebetændelse
75	Pleuritis, Empyema	Lungehindebetændelse
76	Bronchitis acuta simplex	Brystkatarrh, akut bronchitis
77	Bronchitis capill. & Pneumonia catarrh	Kapillær Bronchitis og katarrhalsk Lungebetændelse
78	Bronchit. chron. & Bronchiectasis	Chronisk Bronchitis
79	Emphysema pulmonum & Emphysem, Astma	
80	Alii pulmonum morbi chronici	Andre chroniske Lungesygdomme
81	Peri- & Endocarditis	Betændelse af Hjertet og dets hinder
82	Morbus Cordis	Organisk Hjertesygdom
83	Aneurysma Aortæ	Udvidning af Aorta
84	Phlebitis	Blodaarebetændelse
85	Ulcus perforans ventriculi	Perforerende Mavesaar
86	Peritonitis	Bughindebetændelse
87	Enteritis, Colitis, Typhlitis	Tarmbetændelse
88	Ileus	Tarmslyngning
89	Hernia incarcerata	Indeklemt Brok
90	Cirrhosis hepatis	Lever-Cirrhose
91	Echinococcus hepatis	Lever-Echinokok
92	Cholelithiasis	Galdesten
93	Nephritis albuminosa	Brights Sygdom
94	Lithiasis renalis & vesicalis	Nyre- og Blæresten
95	Cystitis	Urinblærebetændelse
96	Strictura urethræ	Forsnevring af Urinrøret
97	Hypertrophia prostatæ	Chronisk Prostatasygdom
98	Tumor ovarii, Hydrops ovarii	Æggestok-Svulst
99	Alii morbi abdominales chronici	Andre chroniske Underlivssygdomme
100	Hydrops ex ignota causa ortus	Vandsot af ubekjendt Aarsag
101	Alii morbi organorum interiorum	Andre Sygdomme i indvendige Organer
VII	Morbi externarum partium	Sygdomme i de ydre dele
102	Phlegmone, Abscessus	Bindevævsbetændelse
103	Caries & Necrosis ossium	Benedder
104	Arthrocace	Leddebetændelse
105	Fractura coli femoris	Brud af Laarbenets Hals
106	Gangræna	Koldbrand
107	Carbunculus & Furunculus malignus	Brandbyld
108	Alii externarum partium morbi	Andre Sygdomme i de ydre Dele
VIII	Aliæ causæ mortis freqventes	Andre hyppige Dødsårsager
109	Marasmus senilis	Alderdomssvaghed

110	Atrophia infantilis	Tæring hos Smaabørn
111	Mors in partu & puerperio (Fb. puerp. excl.)	Død under Fødslen og i Barselsseng (Barselfeber ikke medregn.)
112	Mors repentina sine nota causa	Pludselig Død uden bekjendt Aarsag
113	Causa mortis vel male vel omnino non indicata	Uangiven eller slet specificeret Dødsaarsag
114	Mors medico non vocato obveniens	Død uden Lægebehandling
IX	Exanimus natus	Dødfødte
115	Exanimis natus	Dødfødt

B. Appendix B: The Heiberg groups

Heiberg group nr.	Heiberg group name	DK-1875 categories
I	Infectious diseases	1-23
II	Croupous Pneumonia	74
III	Tuberculosis	27-31
IV	Cancer	32-35
V	Alcoholism	24-26
VI	Suicide	49-54
VII	Accidents	40-48
VIII	Diseases in the nervous system	61-71
IX	Respiratory diseases	72, 73, 75-80
X	Cardiovascular diseases	81-84
XI	Digestive diseases	85-92
XII	Genitourinary diseases	93-97
XIII	Others	36-39, 55-60, 98-115