



Guide to The Historical Cause of Death Register

Version 1

February 2025

Joen Rommedahl

Amalie Regitze Faber Mygind

Peter Brunsgaard Trolle

Jeppe Klok Due





Indholdsfortegnelse

1 Project and data decription.....	3
1.1 What is HDAR?	3
2 Historical sources	3
2.1 Death certificate formats	4
2.2 Geographical limitations.....	5
2.3 Other know limitations to population	6
2.4 Death certificate information	6
2.4.1 Name	7
2.4.2 Birthdate	8
2.4.3 Death date.....	8
2.4.4 Age	9
2.4.5 Cause of death.....	9
2.4.6 Primary illness	10
2.4.7 Manner of death.....	10
2.4.8 Cause of death code	11
2.4.9 Marital status	12
2.4.10 Occupation.....	12
3 AI methods	13
4 Data structure and processing logic	13
4.1 General processes.....	15
4.2 ID columns	16
4.3 Name columns.....	16
4.4 Dates and age columns.....	18
4.5 Cause of death columns	20
4.6 Marital status and occupation columns	22
4.7 Certificate type and primary source columns	23
5 Evaluation of transcription quality	23
6 Suggested usage and further work.....	40



1 Project and data description

1.1 What is HDAR?

The Historical Cause of Death Register (HDAR) is a by-product of the Multi-Generation Register (MGR) project¹. The MGR project, funded by the Novo Nordisk Foundation, aims to create a digital register containing familial relationships for all individuals born in Denmark from 1920 to the present, using AI-based methods. Once completed, the MGR will enable groundbreaking research in areas such as the heritability of late-onset diseases, social mobility, and other fields where genetics or family structures play a crucial role.

As part of constructing the MGR, it was necessary to determine whether and when individuals had died. Since the first digital registry of deaths in Denmark begins in 1943², identifying deaths before this year required the creation of HDAR. This register was developed using state-of-the-art AI methods in handwritten text recognition. We expect HDAR to be used as a historical extension of the digital cause of death registry, created by the Department of Clinical Epidemiology (DIKE) which spans the years 1943-1969. The existing 1943-1969 registry is in this article referred to as “DIKE-DAR” to distinguish it from HDAR.

Because HDAR is based entirely on publicly available information, we have chosen to make it accessible to the public. We hope it will serve as a valuable resource for anyone interested in medical history, genealogy, or personal historical research.

2 Historical sources

HDAR is based on death certificates sent by the health inspector (embedslæge) to the Danish Health Authority (Sundhedsstyrelsen). Until 1920, official death statistics in Denmark were compiled from information provided by clerical parishes. However, in 1920, this practice changed, and death statistics were thereafter based on medical death certificates.

Due to this shift in record-keeping, nearly all death certificates were sent to the Danish Health Authority and later archived at the Danish National Archives, where they remain today. HDAR is specifically based on the Danish Health Authority's death certificates from 1920 to 1943.

¹ MGR artikel eller lignende

² DIKE DAR reference



2.1 Death certificate formats

During the period covered by HDAR, the Danish Health Authority used several different types of death certificates, each applicable to different situations but containing some shared information.

The death certificate types used to create HDAR are as follows:

Certificate type	Description	Count in HDAR
A	Used for children under the age of 1, who have not died as a consequence of an 'unfortunate event'. Used in the period before 1930. Gets replaced by type A1.	51.926
B	Used for deceased of age 1 or older, who have not died as a consequence of an 'unfortunate event'. Used in the period before 1930. Gets replaced by type B1.	303.014
A1	Used for children under the age of 1, who have not died as a consequence of an 'unfortunate event'. Used in the period from 1930 and onwards. Replaces A.	48.700
B1	Used for deceased of age 1 or older, who have not died as a consequence of an 'unfortunate event'. Used in the period from 1930 and onwards. Replaces B.	408.061
A2	Used for children under the age of 1, who have died as a consequence of an 'unfortunate event', meaning accidents or murder. Used in the entire period.	4.961
B2	Used for deceased of age 1 or older, who have died as a consequence of an 'unfortunate event', meaning accidents, suicide or murder. Used in the entire period.	48.042
C	Used for stillborn children, when a doctor has inspected the body. Used in the entire period, but missing 1932-1937.	15.904
D	Used for children under the age of 1, who have not died as a consequence of an 'unfortunate event'. Filled out by 'lignsynsmænd' (see 2.1.10). Used in the period from 1930 and onwards.	2.801





E Used for deceased of age 1 or older, who have not died as a consequence of an 'unfortunate event', and filled out by 'ligsynsmænd' (see 2.1.10).

12.983

Used in the period from 1930 and onwards.

Ligsynsmand

Used in areas where no medical authority or doctor was present to inspect the body and write the death certificate. In these rural districts, so-called 'ligsynsmænd' were authorized to write death certificates. This is the most basic of the certificate types.

45.826

Used in the entire period but rarely after 1930.

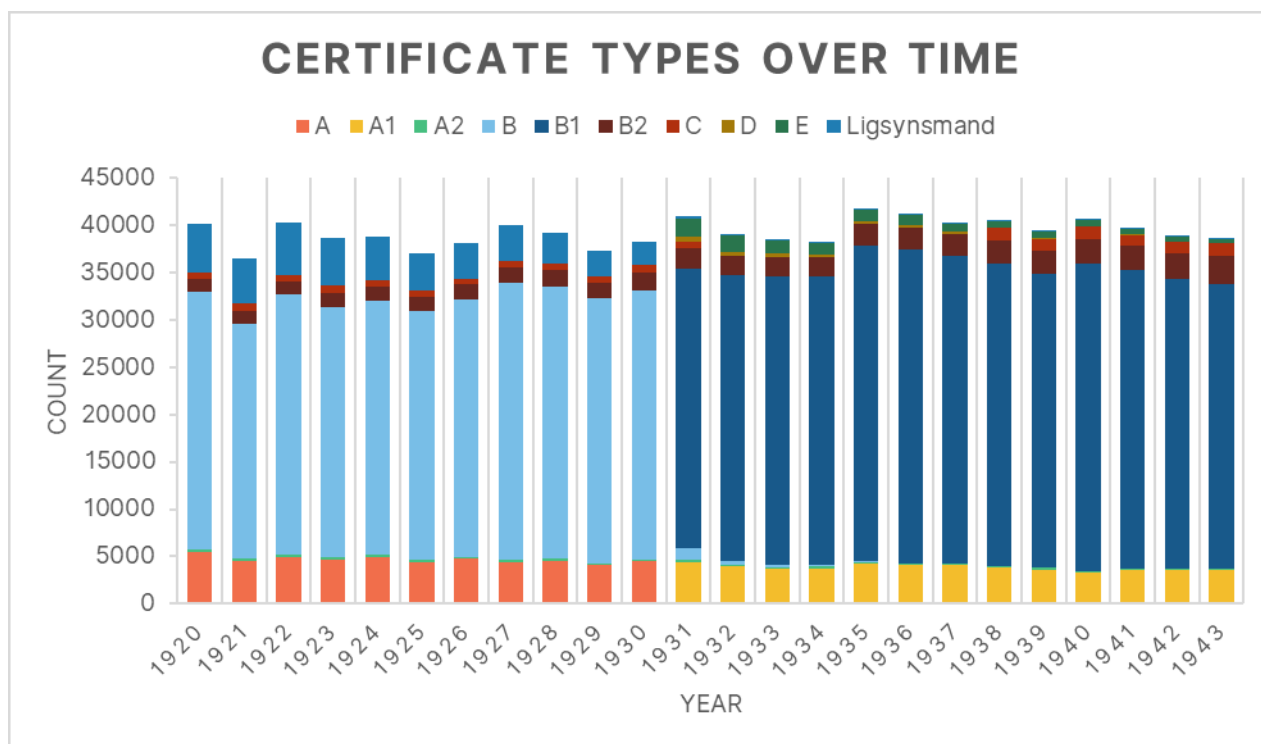


FIGURE 1. DISTRIBUTION OF CERTIFICATE TYPES

Figure 1 shows the distribution of the different certificate types throughout the HDAR period. As evident in the figure, the primary shift in certificate use happens in 1931.

2.2 Geographical limitations

The Danish Health Authority collected death certificates from both national and international sources. However, international death certificates from countries such as Germany, Sweden, and England follow unique formats, making it difficult to develop a



generalized AI model capable of reading them. As a result, these certificates are excluded from HDAR.

Consequently, HDAR is limited to deaths that occurred within Denmark, including Southern Jutland and the Faroe Islands.

2.3 Other known limitations to population

The exclusion of death certificates that deviate from the primary official formats not only means that individuals who died abroad are missing from HDAR but also that stillborn children may be underrepresented. Specifically, stillborn children whose bodies were inspected only by a midwife (rather than a doctor) are not included.

There are no estimates of how many stillborn children are excluded due to this limitation. Additionally, all stillborn deaths (format 'C') are missing for the period 1932-1937, as the corresponding death certificates are absent from the archives for unknown reasons.

2.4 Death certificate information

The information recorded on death certificates varies depending on the format. Not all available information has been extracted for HDAR, and some certificates naturally contain occasional missing data.

Below is an overview of the extracted information, followed by a more detailed description of each data field.

Variable	Formats	Description
Name	A, A1, A2, B, B1, B2, D, E, Ligsynsmand	The name of the deceased
Birthdate	A, A1, A2, B, B1, B2, C, D, E	The birthdate of the deceased
Death date	A, A1, A2, B, B1, B2, C, D, E, Ligsynsmand	The date of death
Age	Ligsynsmand	The deceased's age at death
Cause of death	A, A1, A2, B, B1, B2, C, D, E, Ligsynsmand	The immediate cause of death
Primary illness	A1, B1	The primary illnesses and complications leading to death
Manner of death	A2, B2	The manner of death for those who died of an 'unfortunate event'
Cause of death code	A, A1, A2, B, B1, B2, D, E, Ligsynsmand	The coding of the death certificate to the official statistics





Marital status	B, B1, B2, E, Ligsynsmand	The marital status of the deceased
Occupation	A, A1, A2, B, B1, B2, C, D, E, Ligsynsmand	The occupation of the deceased or his/her parents

2.4.1 Name

1) Fulde Navn (for gift Kvinde, Enke, separeret eller fraskilt tillige Pige-navn). Ugift, gift, Enkemand, Enke, separeret eller fraskilt.	Ane Mathine Petersen f. Poulsen Enke
1) Fulde Navn (for et unavngivet Barn: Køn).	Bent Mathew.
Fulde Navn:	Grinde Hansen

FIGURE 2. EXAMPLES OF NAMES AS WRITTEN ON THE DEATH CERTIFICATES

All death certificate formats, except those for stillborns, contain a field for the deceased's name. In some cases, this field is combined with marital status.

Married women are typically recorded with their maiden name at the end, prefixed by "f.", which stands for "født" (born).

For children who died before receiving a name or for unidentified individuals, various pseudo-names may be used, such as:

- "Udøbt pige" (Unbaptized girl)
- "Ukendt mand" (Unknown man)
- "Havlig" (Drowned person)
- Physical descriptions of the deceased



2.4.2 Birthdate

Fødselsdag og Aar.	
26/8 1848	
Fødselsdag og Aar.	
30-10-1938,	
Fødselsdag:	
d. 18 Jan 19 38 Kl. 7 ⁵⁰	

FIGURE 3. EXAMPLES OF BIRTHDATES AS WRITTEN ON THE DEATH CERTIFICATES

The birthdate is included in all death certificate formats except for the Ligsynsmand format, which records the deceased's age instead.

Birthdates appear in various formats, using dashes, slashes, or dots as separators. Additionally, doctors used numbers, letters, and Roman numerals to record dates.

In many cases, the century is omitted, as doctors assumed it could be inferred from the rest of the document.

2.4.3 Death date

6) Dødsdag.	d. 9/1 19 38 Kl. 1
Dødsdag: 6 Januar	
9) Dødsdag.	d. 31. Oktober 19 38 Kl. 7 ⁴⁵

FIGURE 4. EXAMPLES OF DEATH DATES AS WRITTEN ON THE DEATH CERTIFICATES

The death date is recorded in all death certificate formats except those for stillborns. It follows the same formatting variations as the birthdate.



2.4.4 Age

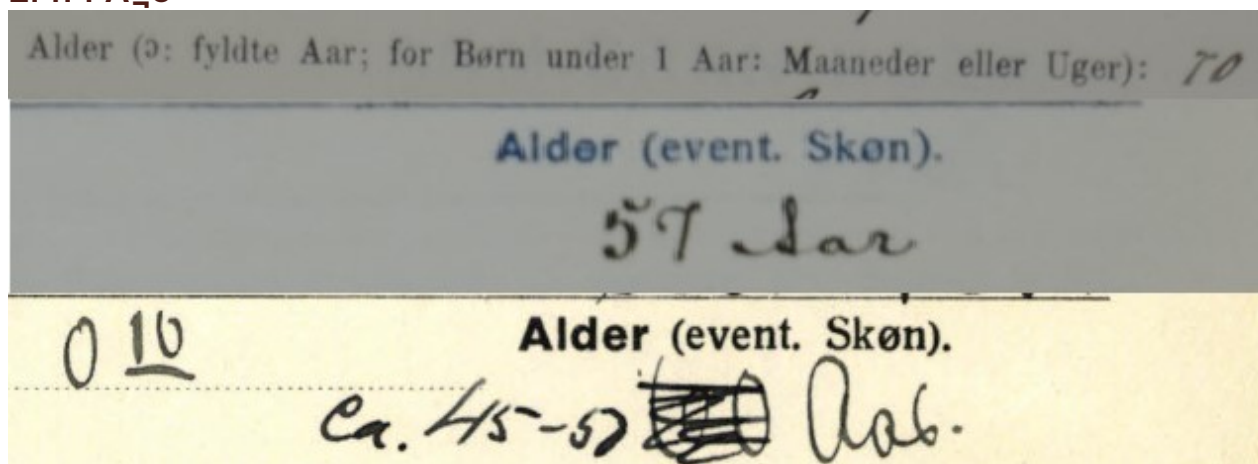


FIGURE 5. EXAMPLES OF AGE AS WRITTEN ON THE DEATH CERTIFICATES

While age is recorded on various certificate formats, it has only been extracted for HDAR from the Ligsynsmand certificates and cases where the deceased was not identified.

In these cases, age is often written as an approximate range, such as:

- "I 30'erne" (In their 30s)
- "25-35" (25 to 35 years old)

2.4.5 Cause of death

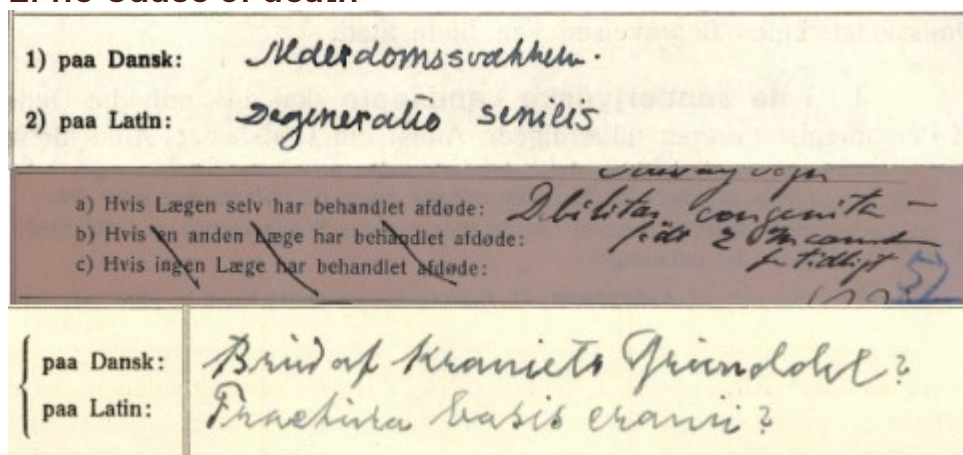


FIGURE 6. EXAMPLES OF CAUSE OF DEATH AS WRITTEN ON THE DEATH CERTIFICATES

The cause of death is recorded on all certificate formats.

- In A and B certificates, both the immediate cause of death and any complications or relevant illnesses are listed.
- From 1930 onward, with the introduction of A1 and B1 certificates, only the immediate cause of death is recorded.

The field is not standardized, meaning a single cause of death may appear with multiple spellings and abbreviations. For example, "Morbus cordis" can be written as:





- "Morbus cordis"
- "Mb. cordis"
- "Mb. cord."
- "Mb. cd."
- ...

For most of the period, causes of death were recorded in both Danish and Latin.

2.4.6 Primary illness

7) Hovedsygdommen..... (paa Latin) og de væsentligste Komplikationer ... (paa Latin).	<i>Tub. plicaria</i> <i>Pleuritis ex. d.</i>
10) Hovedsygdommen..... (paa Latin) og de væsentligste Komplikationer... (paa Latin).	<i>Acute leucipia hatis, genellus I</i> <i>Febus neonatorum, Relictus congenita.</i>

FIGURE 7. EXAMPLES OF PRIMARY ILLNESS AS WRITTEN ON THE DEATH CERTIFICATES

The primary illness field appears only in A1 and B1 formats. In earlier formats (A and B), this information was sometimes written within the cause of death field instead.

The primary illness field follows the same spelling and abbreviation inconsistencies as the cause of death field.

2.4.7 Manner of death

10) Dødsmaaden (naturlig Død, Ulykkestilfælde, Drab.)	Naturlig Død
7) Dødsmaaden. (Naturlig Død, Ulykkestilfælde, Drab, Selvmord).	<i>Ulykkestilfælde</i>
7) Dødsmaaden. (Naturlig Død, Ulykkestilfælde, Drab, Selvmord).	Selvmord

FIGURE 8. EXAMPLES OF MANNER OF DEATH AS WRITTEN ON THE DEATH CERTIFICATES

The manner of death is recorded only in A2 and B2 formats, which apply to deaths caused by unfortunate events.

Ideally, this field should categorize the death as one of the following:

- "Naturlig død" (Natural death)
- "Ulykkestilfælde" (Accident)



- "Mord" (Murder)
- "Selvmord" (Suicide)

However, in many cases, this field contains a longer description of how the deceased died.

2.4.8 Cause of death code

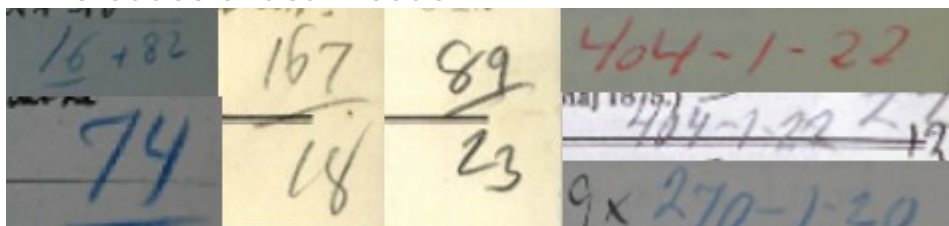


FIGURE 9. EXAMPLES OF CAUSE OF DEATH CODE AS WRITTEN ON THE DEATH CERTIFICATES

Most death certificates have been enriched with various handwritten codes in all likelihood used to create the official death statistics for the time. Some of these codes relate to the cause of death and have been extracted to HDAR. The codes illustrate how the doctors and medical professionals at the Danish Health Authority interpreted the text in the Cause of death, Primary illness and Manner of death fields *at the time*, capturing how the meaning or categorization of certain illnesses have changed over time.

Practically all certificates are given a cause of death code, except for the certificates concerning stillborns.

Identifying the cause of death code on the death certificates is not an easy task as the way of annotating the codes as well as the coding reference system changes across time and geography. According to Juel and Helweg-Larsen³ three reference systems were used in the span of HDAR:

1871-1930

113 codes were used in Denmark in this period. When looking at the actual certificates, it seems that the list expands to about 150 codes in the period ~1928-1930, possibly in preparation for the implementation of the next coding system.

1931-1940

In this period Denmark used a Scandinavian list of causes of death of 199 causes. Sometimes the code written on the certificate refers to the official number of the cause of death in the official list, but in other cases the number refers to a code between 000 and 999, which seems to be another way of referencing these same causes of death by the Danish Health Authority at the time. It has not been possible to find the original mapping table between the number in the official causes of death list and the code used. Instead

³ Knud Juel og Karin Helweg-Larsen: The Danish Register og Causes of Death in Danish Medical Bulletin, vol 46, nr. 4.



we have reverse engineered this bridge table by looking at those examples where certificates were coded after both systems.

1941-1950

In 1941 Denmark adopts the international standard for causes of death with 248 unique causes. All numbers written on the certificates are here a reference to the codes between 000 and 949 the meaning of which have been preserved in the original coding manual.

A complete spreadsheet of the causes of death in HDAR, their numbers, codes and whether they've been found in the official statistics or not can be found in appendix 3.

2.4.9 Marital status

Ugift; gift; Enkemand; Enke: <i>Enke</i>	
<p>1) Fulde Navn, (for gift Kvinde, Enke, separeret eller fraskilt tillige Pigenavn). Ugift, gift, Enkemand, Enke, separeret eller fraskilt. Er afdøde ukendt, oplyses Kønn.</p>	<i>Otto Lindlegaard gift</i>

FIGURE 10. EXAMPLES OF MARITAL STATUS AS WRITTEN ON THE DEATH CERTIFICATES

Marital status is included in formats 'B', 'B1', 'B2', 'E' and 'Ligsynsmand'. In some of these formats the field is shared with the name field. Sometimes the preprinted certificate text has just been underlined. In other cases the marital status is written in the occupation field instead. This is especially true for women, where the occupation field usually specifies their husband's or former husband's occupation, as well as the deceased woman's relationship to him.

2.4.10 Occupation

<p>3) Stilling og Næringsvej (egen, Mandens, Forældrenes, eller Forsørgerens; eventuelt om under offentlig Forsorg eller Aldersrentenyder; for Børn født udenfor Ægteskab Moderens Navn og Stilling). For Børn under 14 Aar: Hos Forældrene eller i Pleje, og da hos hvem?</p>	<p>Gift med Barbermester.</p> <p><i>250</i></p>
<p>5) Faderens (for et udenfor Ægteskab født Barn Moderens) Navn og Næringsvej.</p>	<p><i>081 972 - 1.6.4.58.0.7.36, 081-1-01 961</i></p> <p><i>081 972 - 1.6.4.58.0.7.36, 081-1-01 961</i></p>
<p>3) Stilling og Næringsvej (egen, Mandens, Forældrenes eller Forsørgerens; eventuelt om under offentlig Forsorg eller Aldersrentenyder; for Børn født udenfor Ægteskab Moderens Navn og Stilling). For Børn under 14 Aar: hos Forældrene eller i Pleje, og da hos hvem?</p>	<p><i>E. efter. Hustru.</i></p> <p><i>Aldersrentenyder.</i></p> <p><i>405-1-22</i></p>

FIGURE 11. EXAMPLES OF OCCUPATION AS WRITTEN ON THE DEATH CERTIFICATES



The occupation field exists in some format in all the death certificate formats. When the deceased is a child, the occupation field will usually be the occupation of their primary caregiver and his or her name. For women the field is often used for their husband's occupation and their relationship.

The field is not exclusively used for occupation and will also often contain information about whether the deceased received any monetary support from the government.

3 AI methods

To digitize the handwritten text on death certificates, images underwent a process of classification, segmentation, and handwritten text recognition using state-of-the-art AI models.

For more details on the digitization process, see the document "The AI Processes of HDAR".

4 Data structure and processing logic

The final HDAR data set is a table of 48 columns. This section of the guide will specify what post-processing logic has been applied to the raw HTR-transcriptions in order to create the different fields of HDAR.

The overall modelling of HDAR aims to replicate that of DIKE-DAR as much as possible, in order to make usage of the two registers in conjunction easier.

Column name	Data type	Short description
BilledId	Bigint	Image ID used at the Danish National Archives
PersonId	Bigint	Primary key of the table
FuldNavn	Nvarchar(150)	The raw full name transcription
FuldNavnProb	Float	The probability score of the full name transcription
Fornavn	Nvarchar(150)	Given names of the deceased
Foedenavn	Nvarchar(150)	Surname the deceased was born with
Giftenavn	Nvarchar(150)	Surname acquired through marriage
Kaldenavn	Nvarchar(150)	Nickname of the deceased
NavnEval	Nvarchar(100)	Note regarding processing of name
FuldNavnWC75	Nvarchar(150)	Full name transcription with wildcards replacing low-probability tokens
FuldNavnWC75Prob	Float	The new probability score of the wild-card adjusted full name transcription



Koen	Nchar(1)	The sex of the deceased
Foedselsdato	Nvarchar(50)	The raw birthdate transcription
FoedselsdatoProb	Float	The probability score of the birthdate transcription
Foedselsdato_date	Date	The standardized birthdate of the deceased
FDag	Int	Day part of the birthdate
FMaaned	Int	Month part of the birthdate
FAar	Int	Year part of the birthdate
FoedselsdatoEval	Nvarchar(100)	Notes regarding processing of birthdate
FoedselsdatoWC75	Nvarchar(50)	Birthdate transcription with wildcards replacing low-probability tokens
FoedselsdatoWC75Prob	Float	The new probability score of the wild-card adjusted birthdate transcription
Doedsdato	Nvarchar(50)	The raw death date transcription (with year from metadata)
DoedsdatoProb	Float	The probability score of the death date transcription
Doedsdato_date	Date	The standardized death date of the deceased
DDag	Int	Day part of the death date
DMaaned	Int	Month part of the death date
DAar	Int	Year part of the death date
DoedsdatoEval	Nvarchar(100)	Note regarding processing of death date
Alder	Nvarchar(50)	The standardized age of the deceased
AlderProb	Float	The probability score of the age transcription
AlderFylدتAar	Int	The age of the deceased in years
Doedsaarsag	Nvarchar(800)	The raw cause of death transcription
DoedsaarsagProb	Float	The probability score of the cause of death transcription
Hovedsygdom	Nvarchar(800)	The raw primary illness transcription



HovedsygdomProb	Float	The probability score of the primary illness transcription
Doedsmaade	Nvarchar(550)	The raw manner of death transcription
DoedsmaadeProb	Float	The probability score of the manner of death transcription
DoedsaarsagKode	Nvarchar(50)	The raw cause of death code transcription
DoedsaarsagKodeProb	Float	The probability score of the cause of death code transcription
FK_Doedsaarsag Nomenklatur_DoedsaarsagId	Int	Foreign key to DIM_DoedsaarsagNomenklatur table
Civilstand	Nvarchar(100)	The standardized marital status of the deceased
CivilstandProb	Float	The probability score of the marital status transcription
Erhverv	Nvarchar(1000)	The raw occupation transcription
ErhvervProb	Float	The probability score of the occupation transcription
Attesttype	Nvarchar(50)	Certificate type classification
AttesttypeProb	Float	The probability score of the certificate type classification
PrimaerKilde	Nvarchar(50)	The primary source of the transcriptions
FK_HDAR_Meta_BilledSerield	Bigint	Foreign key to the HDAR_Meta table

4.1 General processes

HDAR is based partly on AI predictions (~80%) and entries made by volunteer genealogists through the National Archives' Crowdsourcing initiative (~20%). Whether a value has been manually entered or automatically predicted can be seen from the associated probability field. A probability value of '-1' indicates that the associated text field has been manually entered. A probability value of '-3' indicates that the associated text string has been calculated based on other information (such as when age is calculated based on date of birth and date of death). Regardless of whether a value has been manually entered or automatically predicted, we refer the the value as having been 'transcribed'.



For some information from the death certificates only the raw transcription is provided, while for others both the raw transcription as well as a processed/standardized value is provided. In a few cases where it made sense, only a standardized value is provided.

4.2 ID columns

There are two ID columns in HDAR: **BilledId** and **PersonId**. The BilledId is an ID created at the Danish National Archives which identifies the image of the death certificate.

Column name	Data type	Short description
BilledId	Bigint	Image ID used at the Danish National Archives
PersonId	Bigint	Primary key of the table

Ideally, only a single death is registered on each certificate. We have however identified a few cases where the doctor have recorded multiple deaths on the same certificate, usually stillborn twins. We therefore couldn't use **BilledId** as the primary key, and instead created the new primary key, **PersonId**, which is calculated as the corresponding **BilledId** * 10 + the sequence number of the deceased on the certificate.

An attempt has been made to find duplicate scans and clean them out, so that there is only one row in HDAR for one death certificate. However, there will undoubtedly still be a few duplicate scans that has not been identified.

It happens in rare cases that a single death is registered on several different death certificates. In these cases, the duplicates are not sorted out, since it has not been possible to identify these cases programmatically. Only a portion of these cases have been removed, where it was possible to identify duplicates by combining the transcriptions of names, birthdates, death dates and other variables.

Duplicates can largely be identified by grouping by **Foedselsdato_date**, **Doedsdato_date** and **FuldNavn**, though that method will also include some false positives. When seeking to identify duplicates, one should also filter out all rows with not null in **FoedselsdatoEval** or **NavnEval**, as well as all where the difference between **Foedselsdato_date** and **Doedsdato_date** is ≤ 1 day, as these are usually unnamed children.

4.3 Name columns

There are 10 columns in HDAR related to the name of the deceased.

Column name	Data type	Short description
-------------	-----------	-------------------



FuldNavn	Nvarchar(150)	The raw full name transcription
FuldNavnProb	Float	The probability score of the full name transcription
Fornavn	Nvarchar(150)	Given names of the deceased
Foedenavn	Nvarchar(150)	Surname the deceased was born with
Giftenavn	Nvarchar(150)	Surname acquired through marriage
Kaldenavn	Nvarchar(150)	Nickname of the deceased
NavnEval	Nvarchar(100)	Note regarding processing of name
FuldNavnWC75	Nvarchar(150)	Full name transcription with wildcards replacing low-probability tokens
FuldNavnWC75Prob	Float	The new probability score of the wild-card adjusted full name transcription
Koen	Nchar(1)	The sex of the deceased

The raw transcription of the name and the probability score thereof are the basis of all subsequent name columns and can be found in **FuldNavn** and **FuldNavnProb**. The transcription model has been trained to try to standardize names into the format of [GivenName] [Surname], f. [Maiden Name], kld. [Nickname] regardless of how it is written on the certificate.

These name parts have been isolated in the columns **Fornavn** (given names), **Foedenavn** (maiden name / the surname the person was born with), **Giftenavn** (surname acquired through marriage) and **Kaldenavn** (nickname). The different name parts have been isolated by using the expected formatting of the raw transcription. This also means, that **Foedenavn** and **Giftenavn** always are just a single word. Any remaining name parts end up in **Fornavn**. Most formatting denotations such as parentheses, commas, "f." and such have been removed from the isolated name part columns. Likewise pseudo-names such as "dødfødt" (stillborn), "udøbt" (unbaptized) and "ukendt" (unknown) have been removed from the isolated name parts. For this reason, the isolated name parts are more suited for linking to other data sets than the raw transcription in **FuldNavn**.

Nicknames are rare but can be both given names, surnames and combinations thereof.

The name field of the death certificates are in some cases used to write things which are not explicit names such as "dødfødt" (stillborn), "udøbt pige" (unbaptized girl) or "ukendt mand" (unknown man). In these cases the corresponding class of "dødfødt", "udøbt" and "ukendt" is noted in the **NavnEval** column. This is to make it easier to filter out these entries as it is assumed they will not be relevant in several HDAR use cases, and when trying to link to other data sets using the name.

In order to make linking with other data sets easier two wild card columns related to the name of the deceased are included in HDAR: **FuldNavnWC75** and **FuldNavnWC75Prob**.



These columns have been created by replacing the tokens of the raw name transcription of probability score below 0,75 with a wildcard character “*”. The raw transcription “Jens Jensen” could look like “Jens Je*sen”, if the model was sufficiently unsure whether the surname was actually “Jessen”. These wild card columns could be useful for linking with other data sets, if no matches are found on the other naming columns. The **FuldNavnWC75Prob** contains the new probability score where the wildcard characters have a probability of 1. Wildcard columns have only been calculated on the portion of data which was not manually annotated and where the death certificate type contained a birthdate.

The column **Koen** evaluates the sex of the deceased. The death certificates have no explicit indication of sex and therefore sex is identified by evaluating the first and second given name against a manually coded list of gendered names. Furthermore the notation of “f.” in the raw name transcription indicates that the deceased was female.

4.4 Dates and age columns

There are 19 columns in HDAR related to dates and ages.

Column name	Data type	Short description
Foedselsdato	Nvarchar(50)	The raw birthdate transcription
FoedselsdatoProb	Float	The probability score of the birthdate transcription
Foedselsdato_date	Date	The standardized birthdate of the deceased
FDag	Int	Day part of the birthdate
FMaaned	Int	Month part of the birthdate
FAar	Int	Year part of the birthdate
FoedselsdatoEval	Nvarchar(100)	Notes regarding processing of birthdate
FoedselsdatoWC75	Nvarchar(50)	Birthdate transcription with wildcards replacing low-probability tokens
FoedselsdatoWC75Prob	Float	The new probability score of the wildcard adjusted birthdate transcription
Doedsdato	Nvarchar(50)	The raw death date transcription (with year from metadata)
DoedsdatoProb	Float	The probability score of the death date transcription



Doedsdato_date	Date	The standardized death date of the deceased
DDag	Int	Day part of the death date
DMaaned	Int	Month part of the death date
DAar	Int	Year part of the death date
DoedsdatoEval	Nvarchar(100)	Note regarding processing of death date
Alder	Nvarchar(50)	The standardized age of the deceased
AlderProb	Float	The probability score of the age transcription
AlderFylدتAar	Int	The age of the deceased in years

The raw transcription of the birthdate and the probability score thereof are the basis of all subsequent birthdate columns and can be found in **Foedselsdato** and **FoedselsdatoProb**.

For the purpose of making sure that the string prediction in **Foedselsdato** actually is a date, a date-column **Foedselsdato_date** has been introduced. The value in **Foedselsdato_date** has been processed in different ways:

- Casting of raw **Foedselsdato** transcription as date value.
- If only a year or a year and month is written on the death certificate, missing date parts are cast as the 1st. "September 1898" will for example be cast as "01-09-1898".
- If an invalid date is read, such as the 31st of February, that day is cast as the 1st of January instead.
- If only an age is written on the death certificate, such as on those of format 'Ligsynsmand', an estimated birthdate is calculated based on the age and the date of death.
 - If the age is written in minutes or hours the birthdate is set to the same as the date of death.
 - If the age is written in days or weeks the birthdate is calculated as the date of death minus the specified age.
 - If the age is written in months the birthdate is calculated as the 1st of the corresponding month.
 - If the age is written in years the birthdate is calculated as the 1st January of the corresponding year.

In general, when missing or invalid values are read from the death certificate the values default to the first of the month in case of missing day-values and January in case of missing month-values. When a birthdate has been calculated on the basis of an age or standardized due to the original transcription being an invalid date it is noted in the **FoedselsdatoEval** column as "Estimeret" or "Estimeret fra alder".



FDag, **FMaaned** and **FAar** isolates respectively the birthday, birthmonth and birtyear of the deceased based on the value in **Foedselsdato_date**.

FoedselsdatoWC75 and **FoedselsdatoWC75Prob** are wildcard columns calculated in the same way as **FuldNavnWC75** and **FuldNavnWC75Prob**. See the section above for an explanation of how they are calculated.

The columns of **Doedsdato**, **DoedsdatoProb**, **Doedsdato_date**, **DDag**, **DMaaned**, **DAar** and **DoedsdatoEval** are created in the same way as their birthdate counterpart. One added process is that for most death dates the year of death is taken from metadata instead of the transcription. This is possible as the Danish Health Authority have ordered the original death certificates by death year. When death year has been taken from metadata this has been noted in **DoedsdatoEval** as "Dødsår fra metadata".

Three columns are related to the age of the deceased: **Alder**, **AlderProb** and **AlderFyldtAar**. **Alder** and **AlderProb** contains the original age transcription from the death certificate and its probability score. Be aware that age has only been read from those death certificates where no birthdate was written. Where an interval of ages are written, a middle value has been chosen. "30-40 år" has been noted as "35 år". In all other cases the age of the deceased noted in **Alder** has been calculated based on the deceased's birthdate and death date and **AlderProb** has been set to '-3'. The unit of the value in **Alder** is either in "minutter" (minutes), "timer" (hours), "dage" (days), "uger" (weeks), "måneder" (months) and "år" (years).

In **AlderFyldtAar** all the ages in **Alder** has been standardized to the latest full year reached. This mostly affects children whose ages have been denoted in days, weeks and months, which are set to zero.

4.5 Cause of death columns

There are 10 columns in HDAR related to the causes of the death.

Column name	Data type	Short description
Doedsaarsag	Nvarchar(800)	The raw cause of death transcription
DoedsaarsagProb	Float	The probability score of the cause of death transcription
Hovedsygdom	Nvarchar(800)	The raw primary illness transcription
HovedsygdomProb	Float	The probability score of the primary illness transcription
Doedsmaade	Nvarchar(550)	The raw manner of death transcription
DoedsmaadeProb	Float	The probability score of the manner of death transcription



DoedsaarsagKode	Nvarchar(50)	The raw cause of death code transcription
DoedsaarsagKodeProb	Float	The probability score of the cause of death code transcription
FK_DoedsaarsagNomenklatur_DoedsaarsagId	Int	Foreign key to DIM_DoedsaarsagNomenklatur table

Most of these columns have not undergone any form of processing. **Doedsaarsag**, **DoedsaarsagProb**, **Hovedsygdom**, **HovedsygdomProb**, **Doedsmaade** and **DoedsmaadeProb** all contain the raw transcription of the corresponding field and the probability score of the transcription.

DoedaarsagKode and **DoedsaarsagKodeProb** contain the raw transcription of the cause of death code found on the certificate. If no code is found, the value is set to "nan" (for 'not a number'). If multiple death codes are found on the death certificate, as is often the case before 1930 and after 1940, only the first value has been transcribed as it is assumed to be the primary cause of death. **FK_DoedsaarsagNomenklatur_DoedsaarsagId** contains a foreign key to the DIM_DoedsaarsagNomenklatur table. The DIM_DoedsaarsagNomenklatur table contains a master list of causes of death as found in historical editions of the publication "Dødsaarsager i Kongeriget Danmark"⁴ by the Danish Health Authority. Linking the individual death certificate to the master list of causes of deaths was carried out through the following, rather complex process.

1. If the year of death (**DAar**) is before 1931:
 - a. Match the transcribed cause of death code (**DoedsaarsagKode**) with the statistical code used in the period.
 - b. If the transcribed code is among the unofficial, expanded list of codes used ~1928-1930, and the year of death is 1928-1930, they are matched with their corresponding cause of death in the official list, using a manual bridging table created by the project participants themselves. For example, the code '133', which seems to be detailed cancer recti or cancer coli, is matched with the code '33' which was the official code for "Other types of cancers" as only cancer ventriculi, uteri and mammae had their own explicit code in the official list of causes of deaths 1871-1930. This bridging table can be found in appendix 2.
2. If the year of death is between 1931 and 1940:
 - a. If the year of death is before 1937 and the certificate is from either Copenhagen or the islands (using metadata in the tabel HDAR_Meta) and the transcribed code was written in the upper right corner as a fraction...
 - i. The transcribed code is matched with the statistical code used in the period.

⁴ Reference til publikation





- b. If the certificate is from Copenhagen after 1936 or from the rest of the country...
 - i. The transcribed code is matched with codes used by the Danish Health Authority in the period.
 3. If the year of death is after 1940:
 - a. The transcribed code is matched with codes used by the Danish Health Authority in the period.

If after this process no link has been established, the certificate is linked with value '-1' meaning "unknown code". All certificates where no cause of death code has been found is linked to entry '-3'. Please note that linking with the official lists of causes of deaths are done through year of death, geographical location, location of code on the certificate and the code transcriptions alone. The free text transcriptions of cause of death, primary illness and manner of death are not used for neither linking nor verification, as this has been outside the scope of this project.

4.6 Marital status and occupation columns

There are 4 columns in HDAR related to marital status and occupation.

Column name	Data type	Short description
Civilstand	Nvarchar(100)	The standardized marital status of the deceased
CivilstandProb	Float	The probability score of the marital status transcription
Erhverv	Nvarchar(1000)	The raw occupation transcription
ErhvervProb	Float	The probability score of the occupation transcription

Civilstand contains the marital status of the deceased person, standardized to

- "Ugift" (unmarried)
- "Gift" (married)
- "Enke" (widowed)
- "Separeret" (separated)
- "Fraskilt" (divorced)
- "Uoplyst" (unknown)

The raw transcription of marital status is not saved in HDAR. Young children and stillborn are classified as unmarried through use of the certificate types. The occupation transcription is also used in determining the marital status as the occupation "E.e. Gaardmand" means "widow after farmer", hence classifying the individual as widowed. **CivilstandProb** contains the probability score for the raw marital status transcription.



The **Erhverv** and **ErhvervProb** columns contain the raw occupation transcription as well as the corresponding probability score. If the model has read no value from the occupation field, the value in **Erhverv** is set to "Uoplyst" (undisclosed). For young children and stillborn – those with certificate types A, A1, A2, C and D – the occupation transcription is manually prefixed with a "S.a." or "D.a." meaning 'son of' and 'daughter of' as the transcription is assumed to be the parent's occupation.

4.7 Certificate type and primary source columns

These last 4 columns contain mostly meta data and is included in order to help enable researchers further developing HDAR outside the National Archives.

Column name	Data type	Short description
Attesttype	Nvarchar(50)	Certificate type classification
AttesttypeProb	Float	The probability score of the certificate type classification
PrimaerKilde	Nvarchar(50)	The primary source of the transcriptions
FK_HDAR_Meta	Int	Foreign key to the HDAR_Meta table

Attesttype and **AttesttypeProb** contain the classification of the death certificate type and the probability score of the classification models prediction.

PrimaerKilde shows whether the certificate was primarily manually annotated by crowdsourcing (~20%) or transcribed with AI (~80%). Whether a specific value has been manually annotated or not can also be seen by the corresponding probability score being '-1'.

FK_HDAR_Meta is a foreign key to the HDAR_Meta table which contains metadata from the National Archives. Specifically the year and geographical location of the death have been used in processing death date and cause of death linkage.

5 Evaluation of transcription quality

Besides the Character Error rate (CER) of the individual models listed in the "The AI Processes of HDAR" document, the quality of certain variables in HDAR has been evaluated in two ways:

First by comparing with a manually annotated verification data set of ~300 randomly selected death certificates. Not all variables have been annotated for all ~300 certificates.





Secondly by comparing aggregate numbers from HDAR with the official published death statistics for the period.

5.1 Evaluation of selected variables

Name

309 names have been annotated in the verification data set.

- 72.8% of full names were read 100% correctly
- Among those whose full name was not read 100% correctly, the average levenshtein distance was 4.4

Some of the major levenshtein distances come from cases where the order of the name parts are not the same in the annotated and AI-transcribed data sets. When comparing “Albert Christian Louis Jensen” with “Albert Louis Christian Jensen” for example, we get a levenshtein score of 12, even though the individual name parts are correctly transcribed.

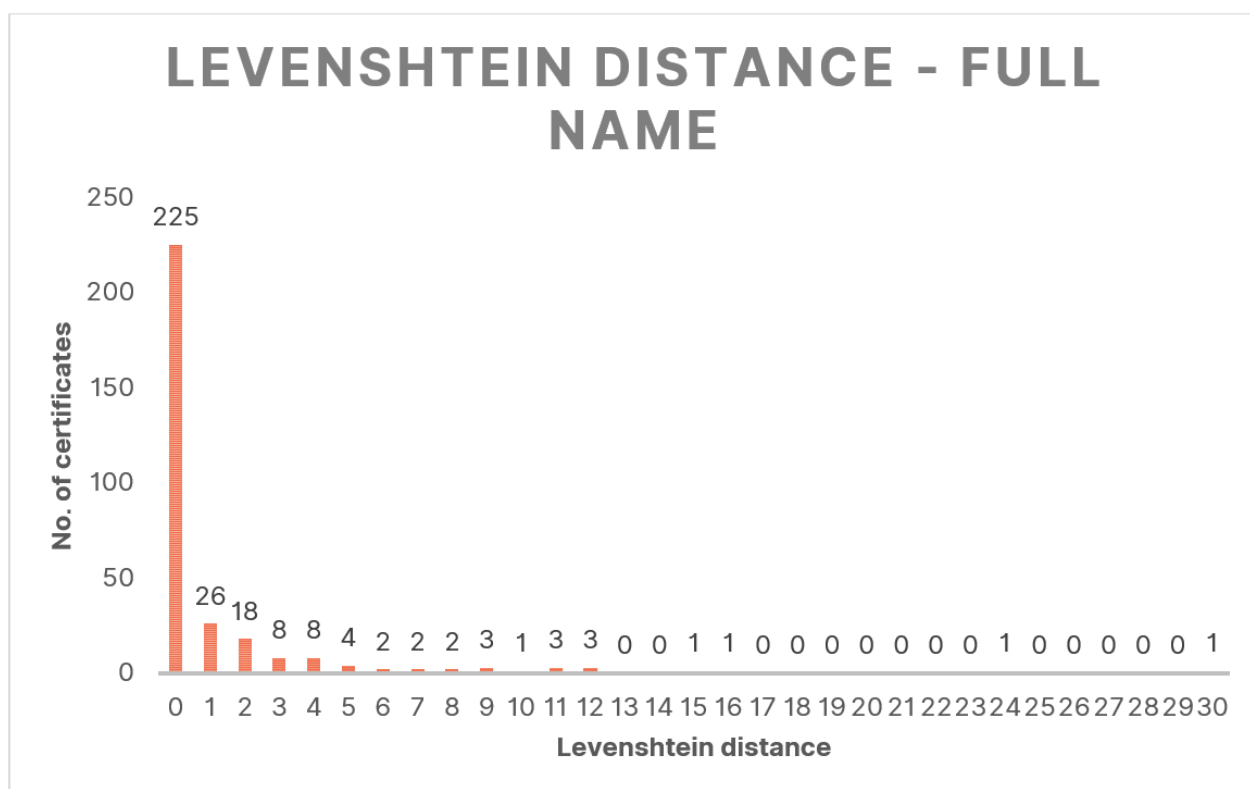


FIGURE 12. LEVENSHTEIN COMPARISON WITH VERIFICATION SET. NAME

Date of birth

285 birth dates have been annotated in the verification data set, as certificate types without an explicit birth date has been excluded.





- 96,5% of birth dates were read 100% correctly
- Among those whose birth date was not read 100% correctly, the levenshtein distance was 1 in all 10 cases

The levenshtein distance illustrates, that when the transcription of birth date is wrong, it is usually only a single digit. This might be useful in linking to other data sets. In these cases, better linkage might be achieved by used the FoedselsdatoWC75 field, when available.

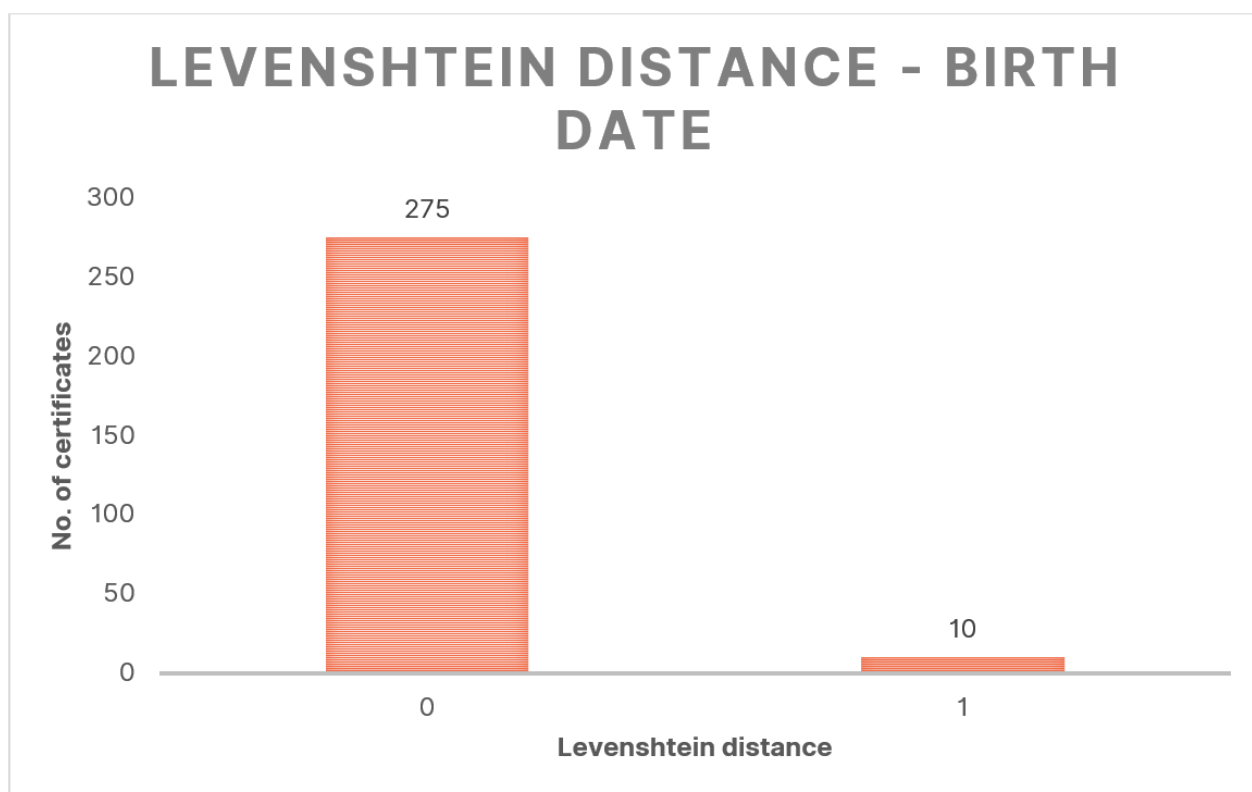


FIGURE 13. LEVENSHTEIN COMPARISON WITH VERIFICATION SET. BIRTH DATE

Date of death

309 death dates have been annotated in the verification data set.

- 96,1% of death dates were read 100% correctly
- Among those whose death date was not read 100% correctly, the levenshtein distance was 1 in all 12 cases

The levenshtein distance illustrates, that when the transcription of death date is wrong, it is usually only a single digit.



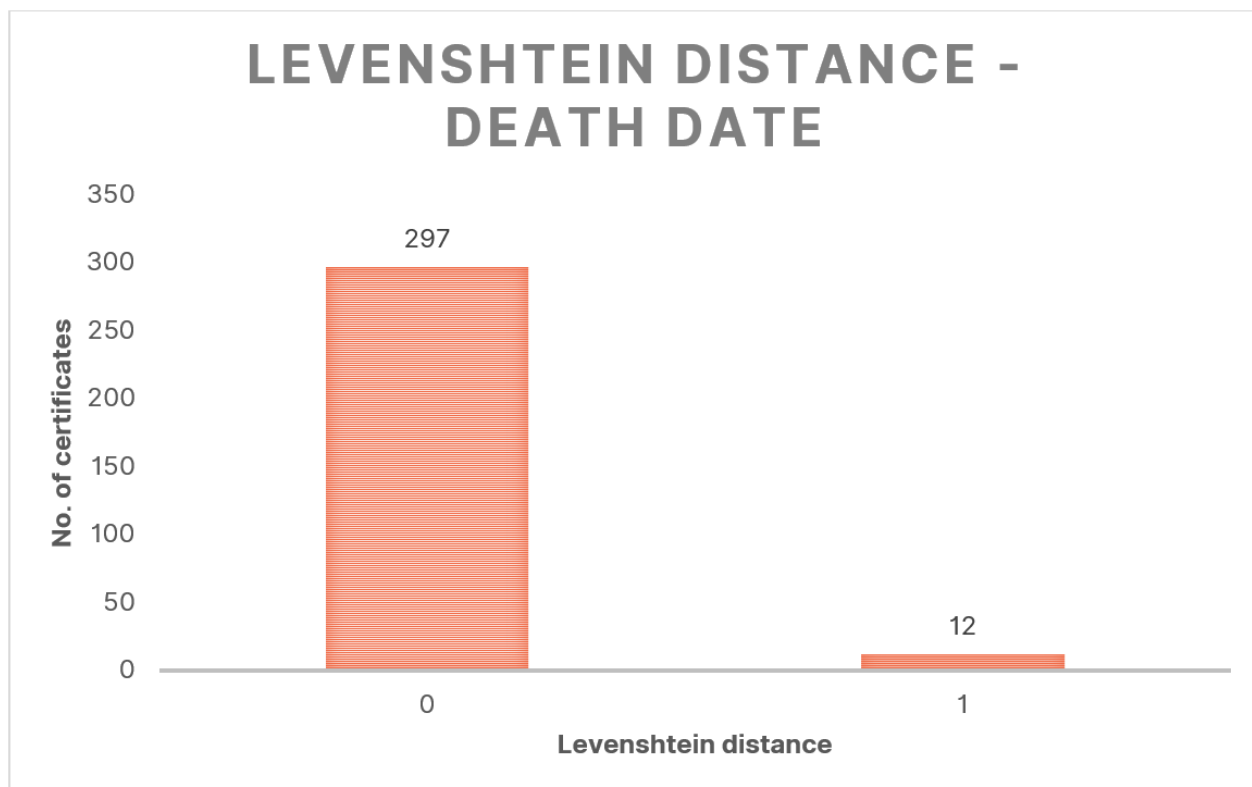


FIGURE 14. LEVENSHTEIN COMPARISON WITH VERIFICATION SET. DEATH DATE

Cause of death

Evaluating cause of death has been very difficult, as our manual annotator had significant difficulties reading the actual handwriting on the certificates. Thus only 40 certificates have been annotated with cause of death in the verification data set.

- 82.5% of cause of death were read 100% correctly
- Among those whose cause of death was not read 100% correctly, the average levenshtein distance was 4.3

Due to the very small sample size however, the actual quality of the cause of death transcriptions is yet unknown. Furthermore, evaluating the free text is difficult as the transformer architecture of the AI-model might in some cases standardize the transcriptions, such that “tub. pulm.” for example is transcribed as “tub. pulmonum”. A proper evaluation of these fields should therefore involve some explicit standardization and grouping of causes of death, to compare the values fairly.



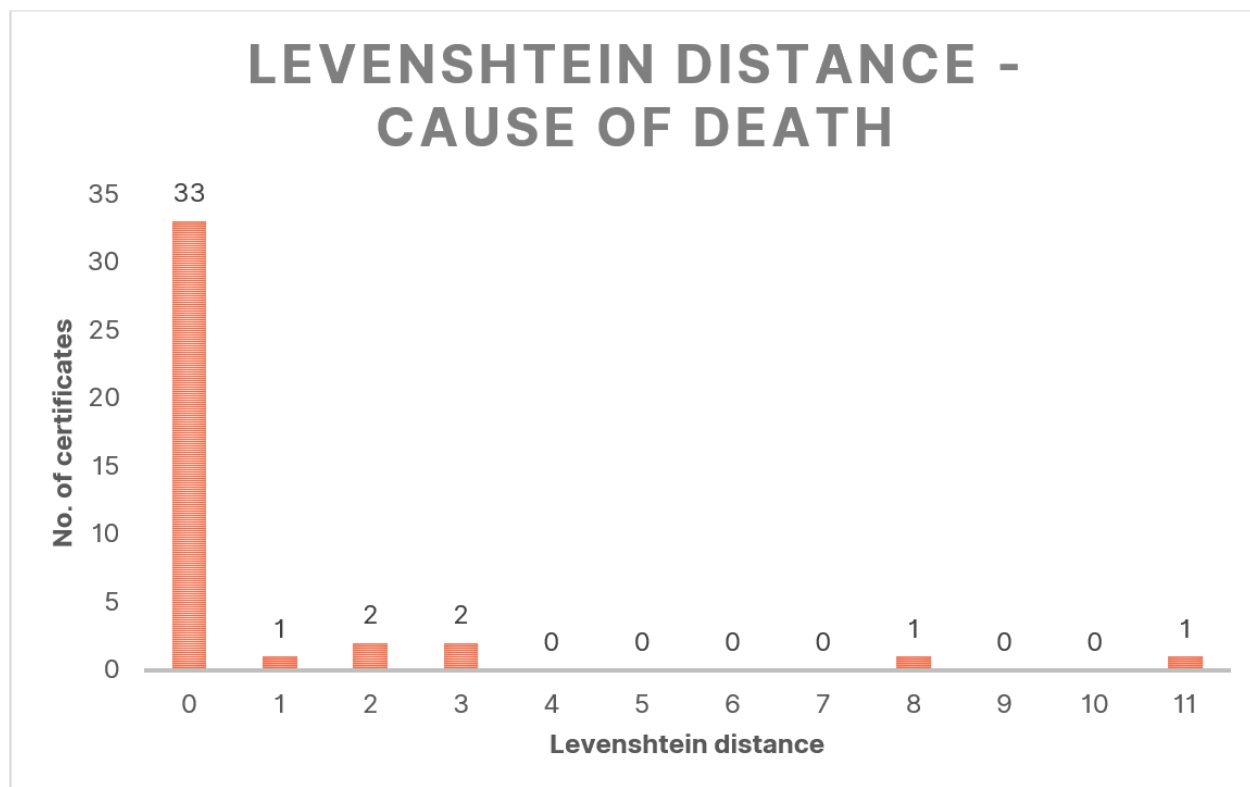


FIGURE 15. LEVENSHTEIN COMPARISON WITH VERIFICATION SET. CAUSE OF DEATH

Cause of death code

299 cause of death codes have been annotated in the verification data set, as certificate types without a cause of death code have been excluded.

- 98.0% of cause of death codes were read 100% correctly
- Among those whose cause of death was not read 100% correctly, the average levenshtein distance was 1.3

If further work is done to verify the death codes, the fact that some digits are correctly transcribed even in most erroneous transcriptions might be useful.



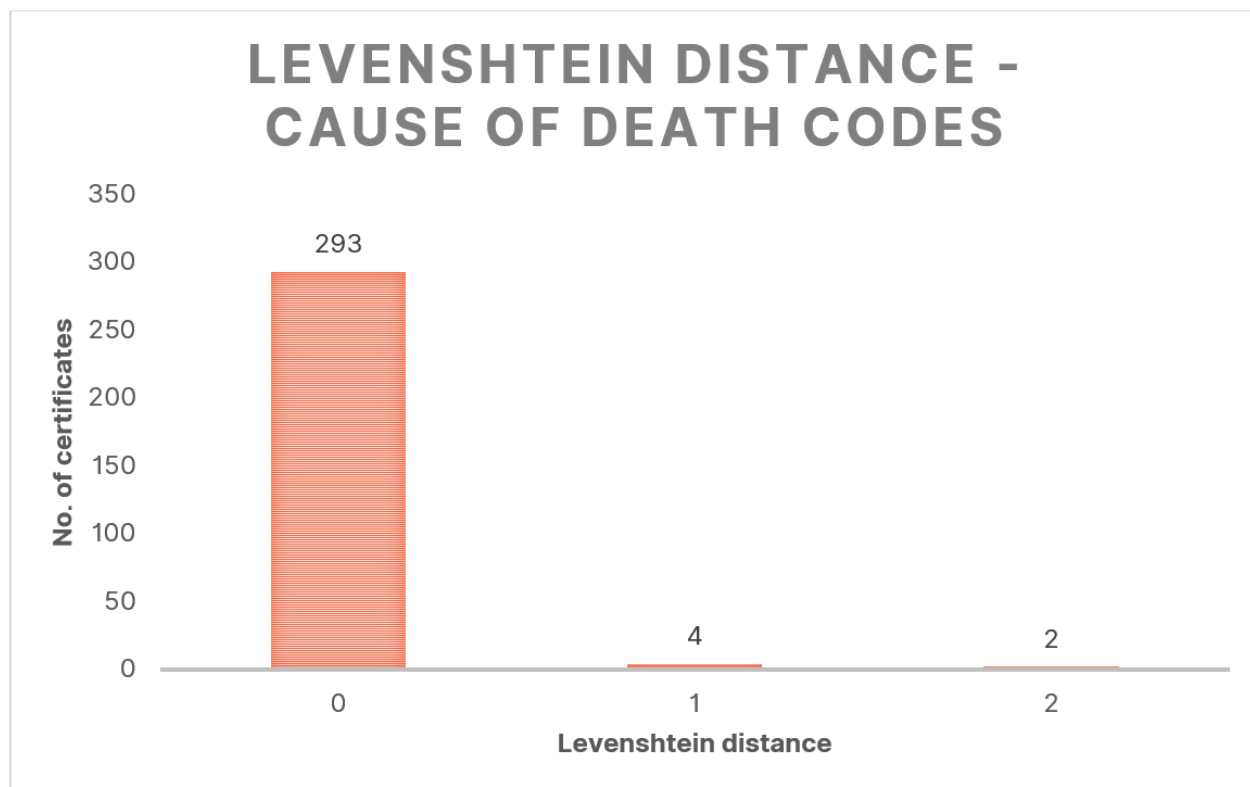


FIGURE 16. LEVENSHTEIN COMPARISON WITH VERIFICATION SET. CAUSE OF DEATH CODE

Marital status

309 marital statuses have been annotated in the verification data set.

- 95,5% of marital status's were classified correctly

As the marital status has been standardized to one of 6 values, levenshtein distance is not an appropriate metric to use for evaluating quality. Instead, a confusion matrix is included below.

Actual\predicted	Ugift	Gift	Separeret	Fraskilt	Enke	Uoplyst
Ugift	116	1	0	0	1	0
Gift	2	105	0	0	0	0
Separeret	0	0	1	0	0	0
Fraskilt	0	1	0	2	1	0
Enke	2	2	0	0	71	0
Uoplyst	2	0	0	1	1	0

The model seems to have difficulties assigning the 'Uoplyst' (unknown) value. This is probably due to errors in training data, as training data is created by genealogists who





often had more knowledge about the deceased individual and has inferred the true value even though it is not present in the certificate. Combined with the fact that many doctors apparently decided not to write the marital status of the deceased, it has probably prompted the annotators of the training data to guess in a lot of cases, rather than write 'Uoplyst'.

Both 'Separeret' and 'Fraskilt' are such rare values, that the quality of the transcription is hard to judge from the 309 verification samples. However it does seem like the model also is significantly worse at classifying 'Fraskilt' correctly – here only 50% is classified correctly – than the more common values.

The values of the **Civilstand** columns are hence more reliable in differentiating between those who are either unmarried, married or widowed, which is most of the population.

Occupation

309 occupations have been annotated in the verification data set.

- 58.6% of occupations were read 100% correctly
- Among those whose occupation was not read 100% correctly, the average levenshtein distance was 9.6

Just as with causes of death, the evaluation of occupations is made difficult by the random and inconsistent use of abbreviations in both model output and the words on the certificate.

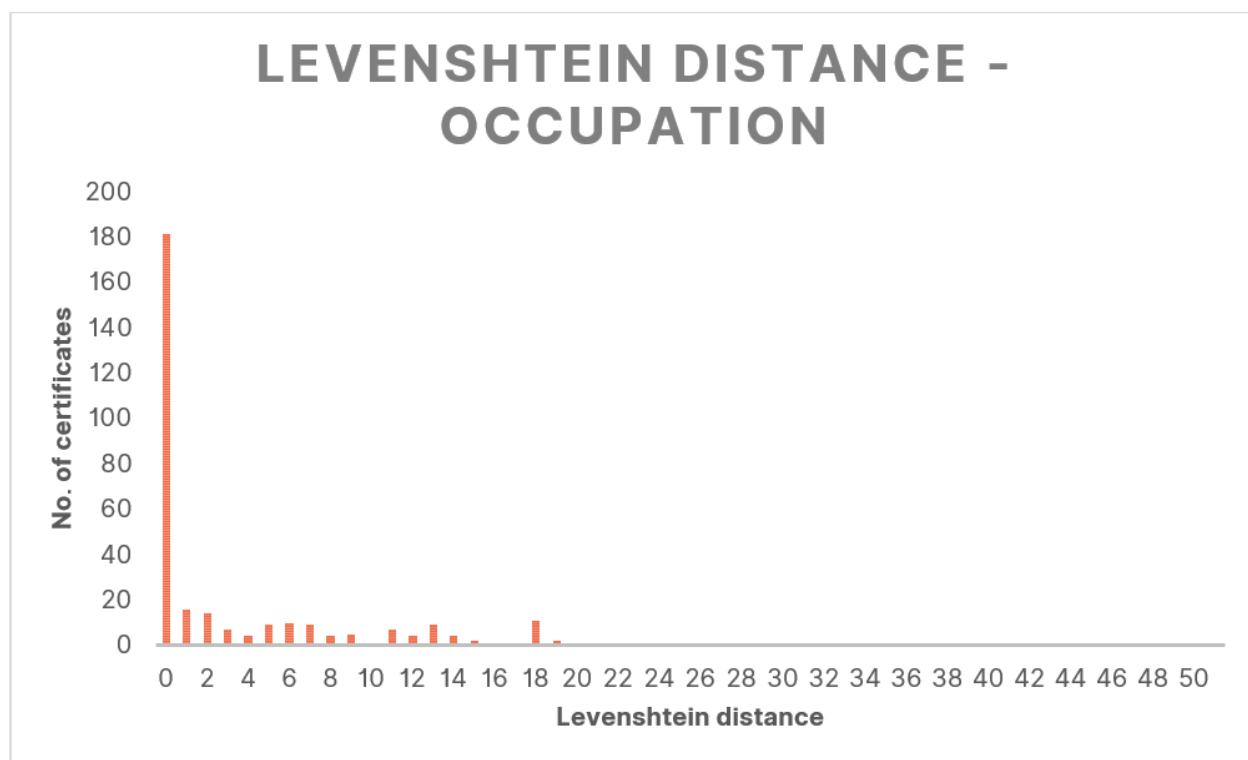


FIGURE 17. LEVENSHTEIN COMPARISON WITH VERIFICATION SET. OCCUPATION



Perhaps even more consequential for the quality of the occupation prediction is that for almost all death certificate types, the field potentially contains much more information than just the occupation. The field can also contain the deceased's partner's or parent's names as well as notes, that are not in itself occupations, but details about the the deceased's economic situation, such as being a housewife or receiving public financial aid. For all of these variations it has been up to the training data annotator to judge, what information is or isn't a part of the occupation. This has undoubtedly led to inconsistent annotations making it more difficult for the model to evaluate what should be considered relevant information and what should be considered noise. It is assumed, that besides giving occasional predictions that are just flat out wrong, this also leads to the model sometimes ignoring some of the extra information that is written in the field, that might be relevant. This can be illustrated by looking at the difference in numbers of words read for the mismatching cases.

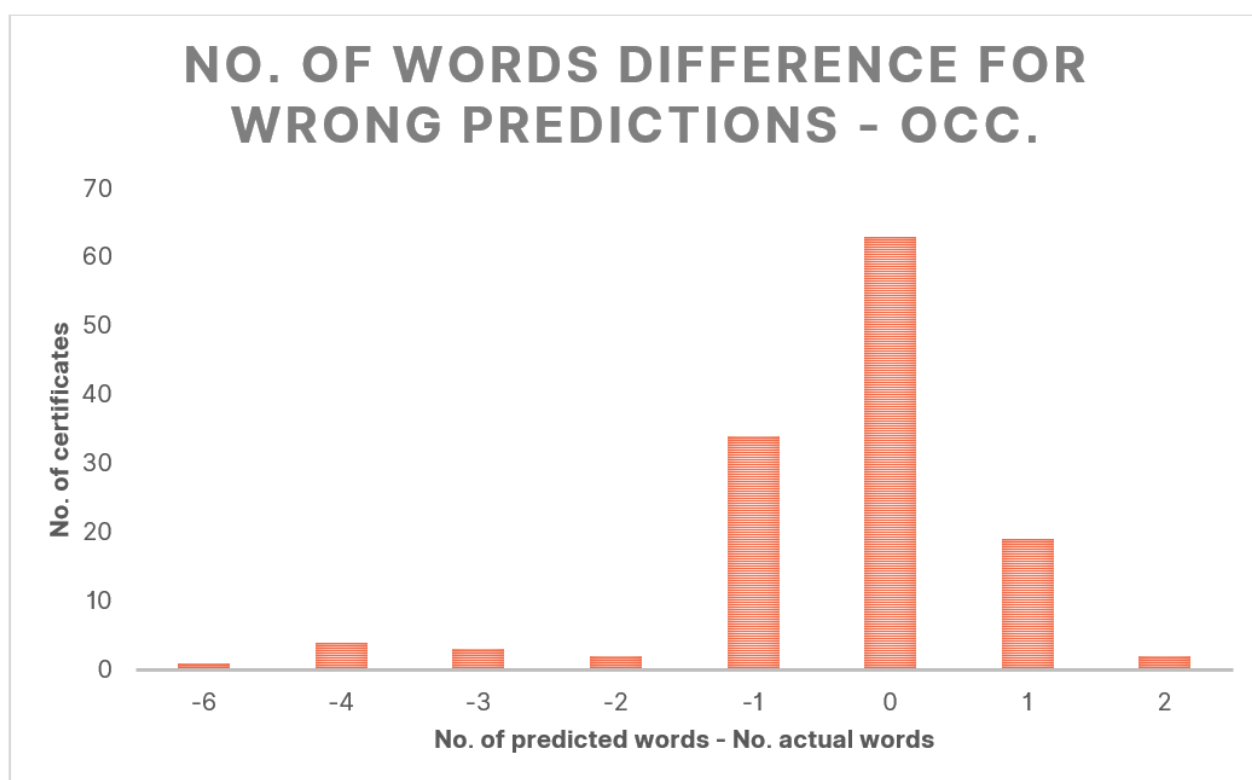


FIGURE 18. NO. OF WORDS PREDICTED COMPARISON WITH VERIFICATION SET. OCCUPATION

The above diagram only shows the cases where the levenshtein distance was not equal to zero. It illustrates the propensity of the model to – when it makes erroneous predictions – misjudge how much of what is written in the field is relevant to the occupation. This can for example lead to the text “Husmoder, g.m. arbejdsmand” to be read as “Husmoder” or “Rentier, fhv. Gaardejer” as “Fhv. gårdejer”. Both examples are from the verification data set.



5.2 Evaluation through statistical comparisons

There are several possible ways to compare HDAR with the official death statistics. In this section, we will look at how closely the cause of death distribution in HDAR aligns with the corresponding distributions in the official statistics.

5.2.1 Comparison with DIKE-DAR

The year 1943 is covered by both HDAR and DIKE-DAR, which makes it possible to compare to two data sets directly.

In terms of overall numbers, HDAR contains more individuals with 38.591 deceased while DIKE-DAR has 37.883 deceased for 1943. Some of this might be the previously mentioned duplicates, but most are expected to be stillborn children (certificate type 'C') which it seems was not or perhaps only partially included in DIKE-DAR.

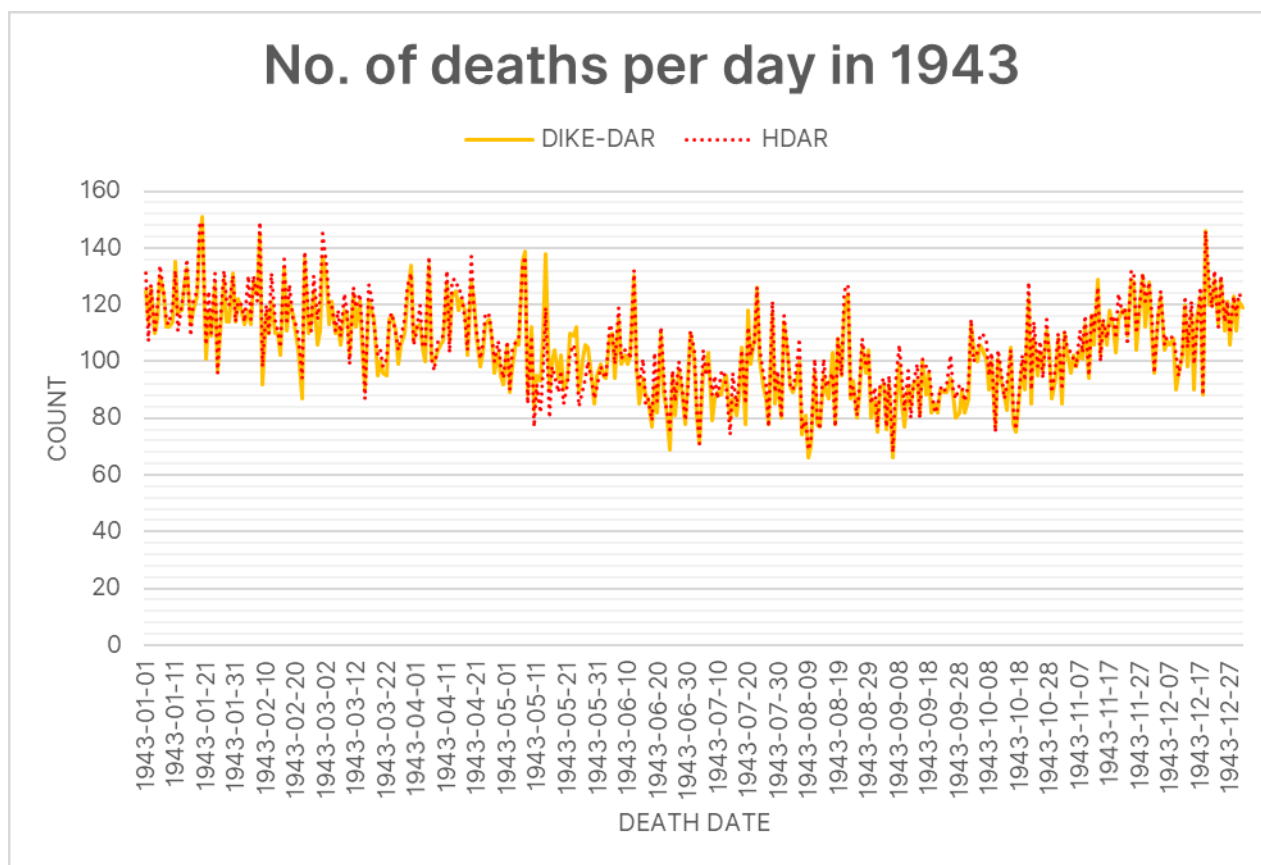


FIGURE 19. COMPARISON DIKE-DAR AND HDAR. NO. OF DEATHS

The graph above compares how many deaths per day in 1943 are registered in HDAR and DIKE-DAR. As can be seen visually, the overall trends are aligned very closely, albeit small differences do exist. The average absolute difference between number of deaths on a single day between the two data sets is 3,68. The median absolute difference is 3.





Below is a series figures showing the birth year distribution for individuals who have died in 1943, 1942, 1940, 1930 and 1920. For 1943, both DIKE-DAR and HDAR is represented in the figure for 1943.

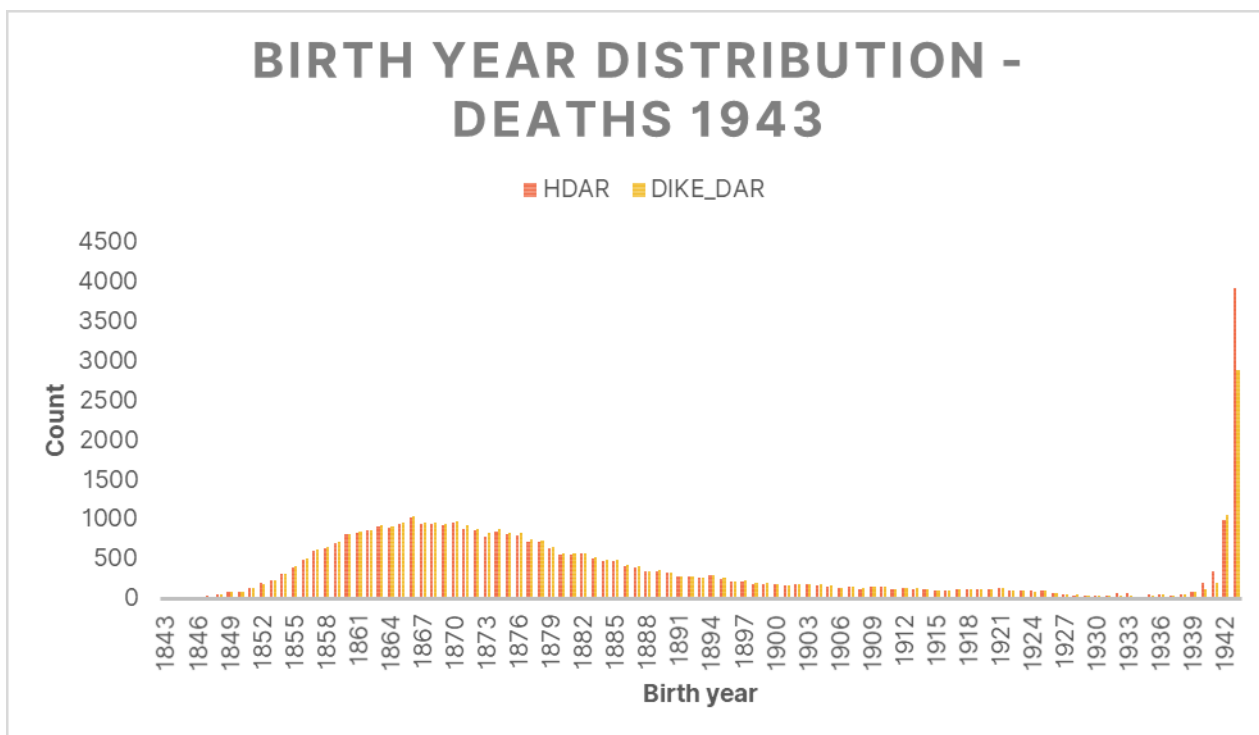


FIGURE 20. COMPARISON DIKE-DAR AND HDAR. BIRTH YEAR FOR DEATHS IN 1943



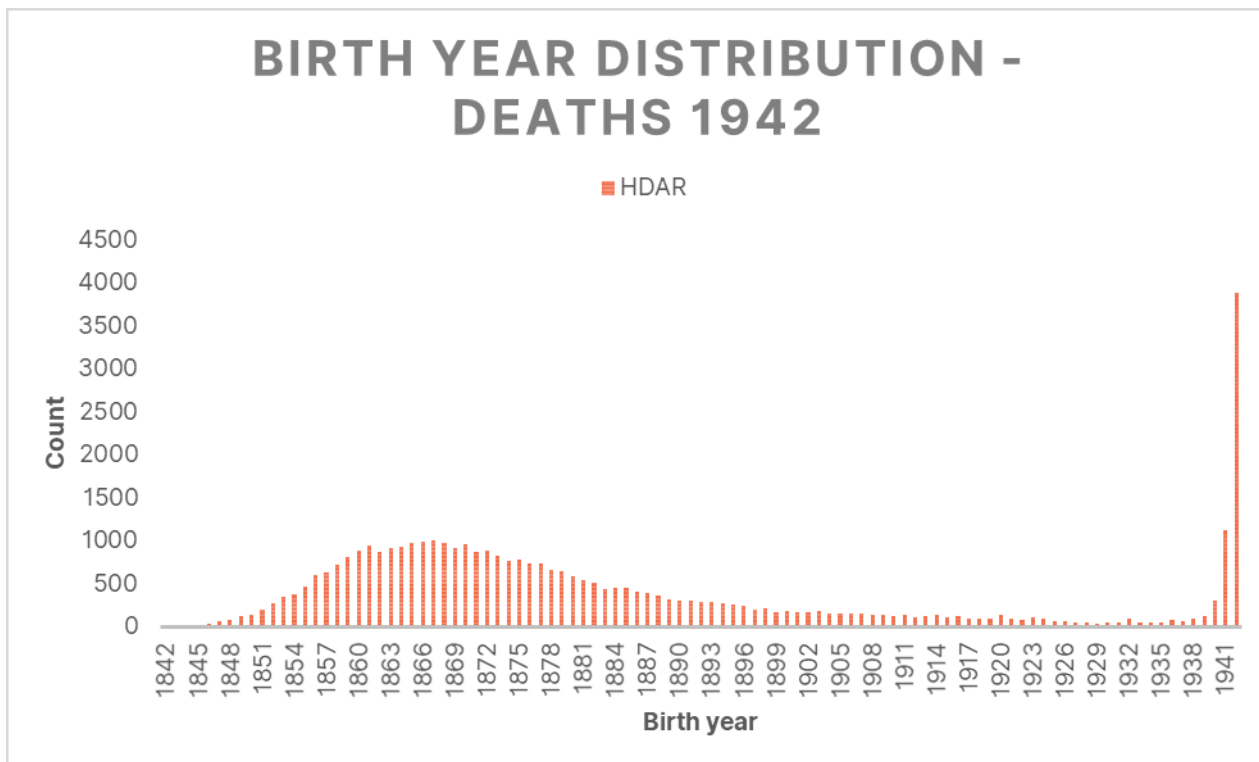


FIGURE 21. BIRTH YEAR FOR DEATHS IN 1942

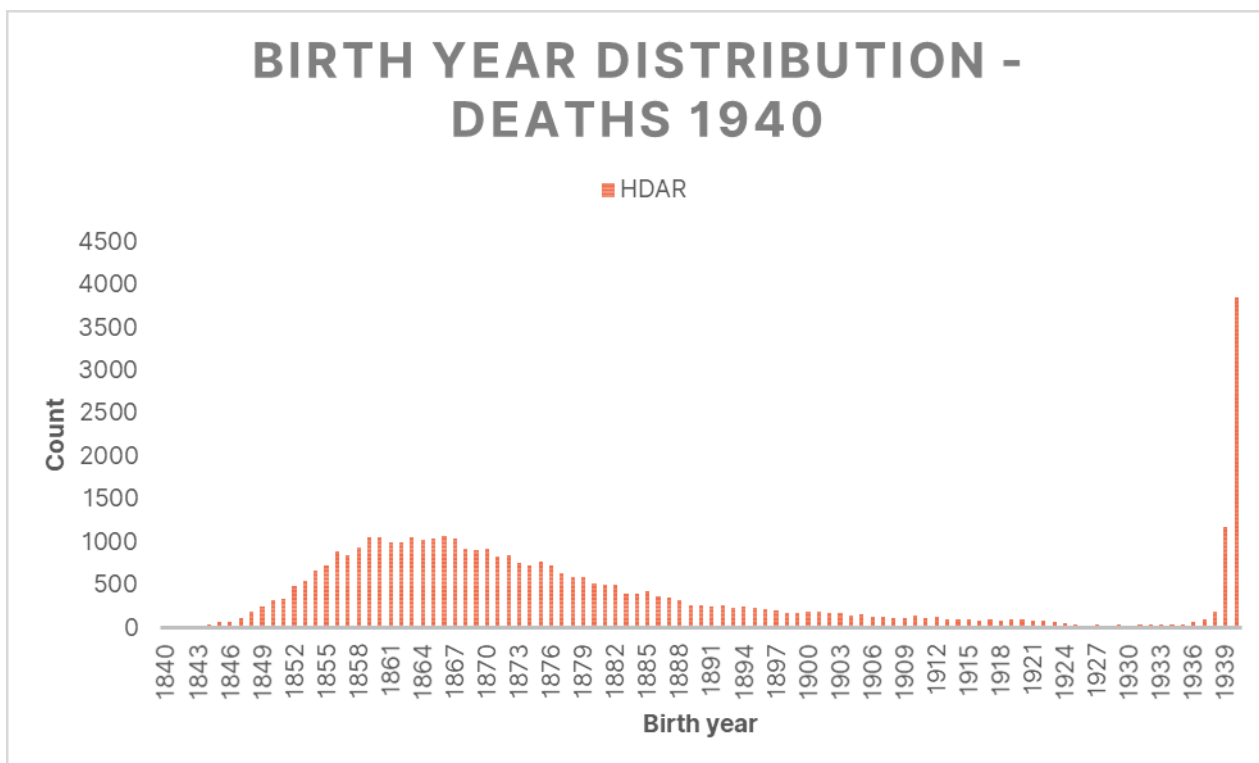


FIGURE 22. BIRTH YEAR FOR DEATHS IN 1940



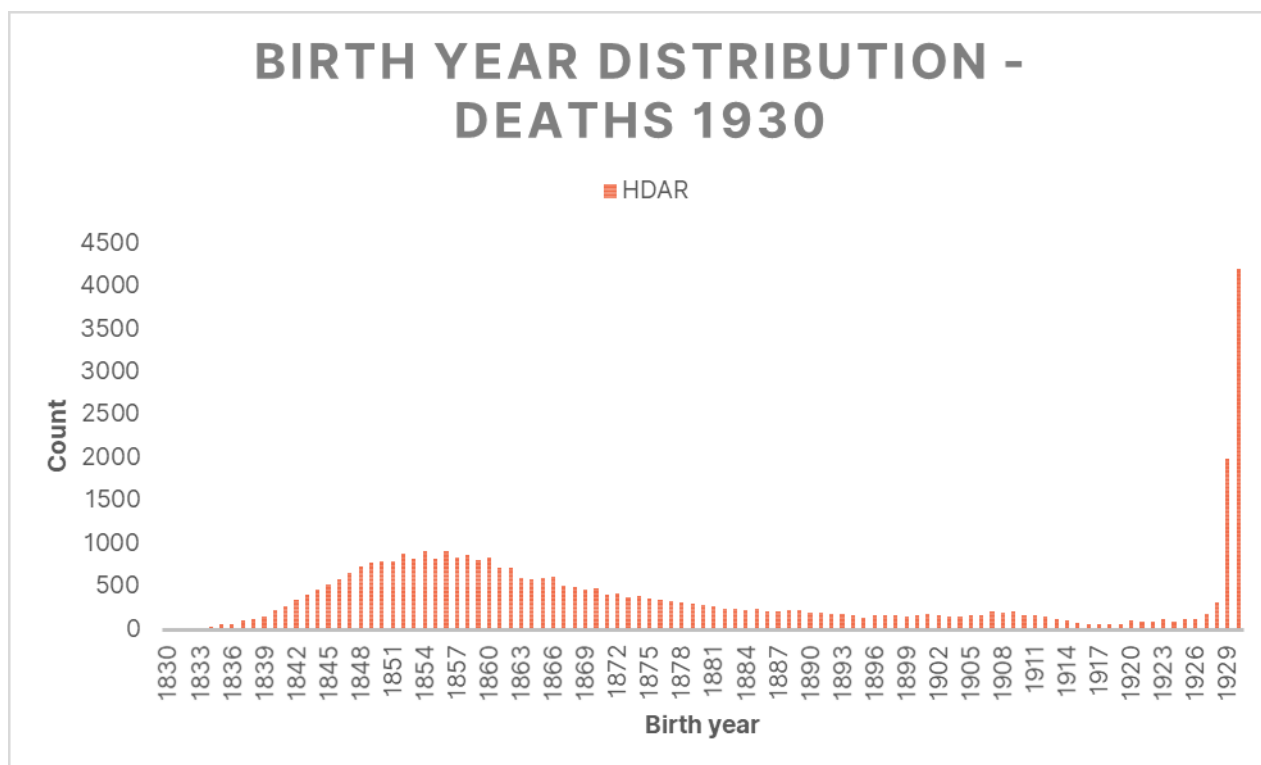


FIGURE 23. BIRTH YEAR FOR DEATHS IN 1930

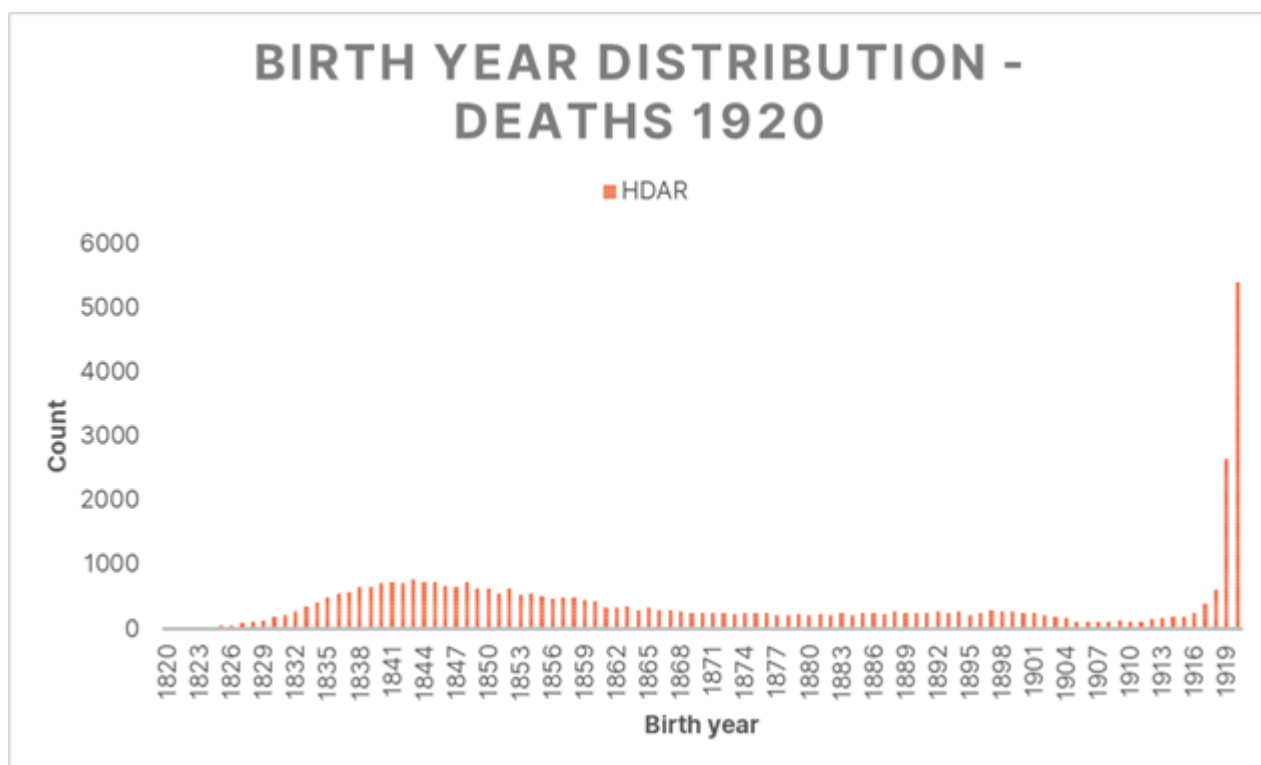


FIGURE 24. BIRTH YEAR FOR DEATHS IN 1920



As seen in the figure for 1943, the birth year distribution largely follows the same patterns between DIKE-DAR and HDAR, except for children under the age of 1. There are many deceased under the age of 3, followed by a fairly constant number of around a few hundred deceased of every age until people get in their 60'ies to 80'ies, of which a larger number are dying. This same pattern is repeated in 1942, 1940, 1930 and 1920, except many more children under the age of 1 is registered in 1920 compared to the other year analyzed. We assume this is due to the general development of the medical community in Denmark and the progressive lowering of childrens mortality throughout the 20th century.

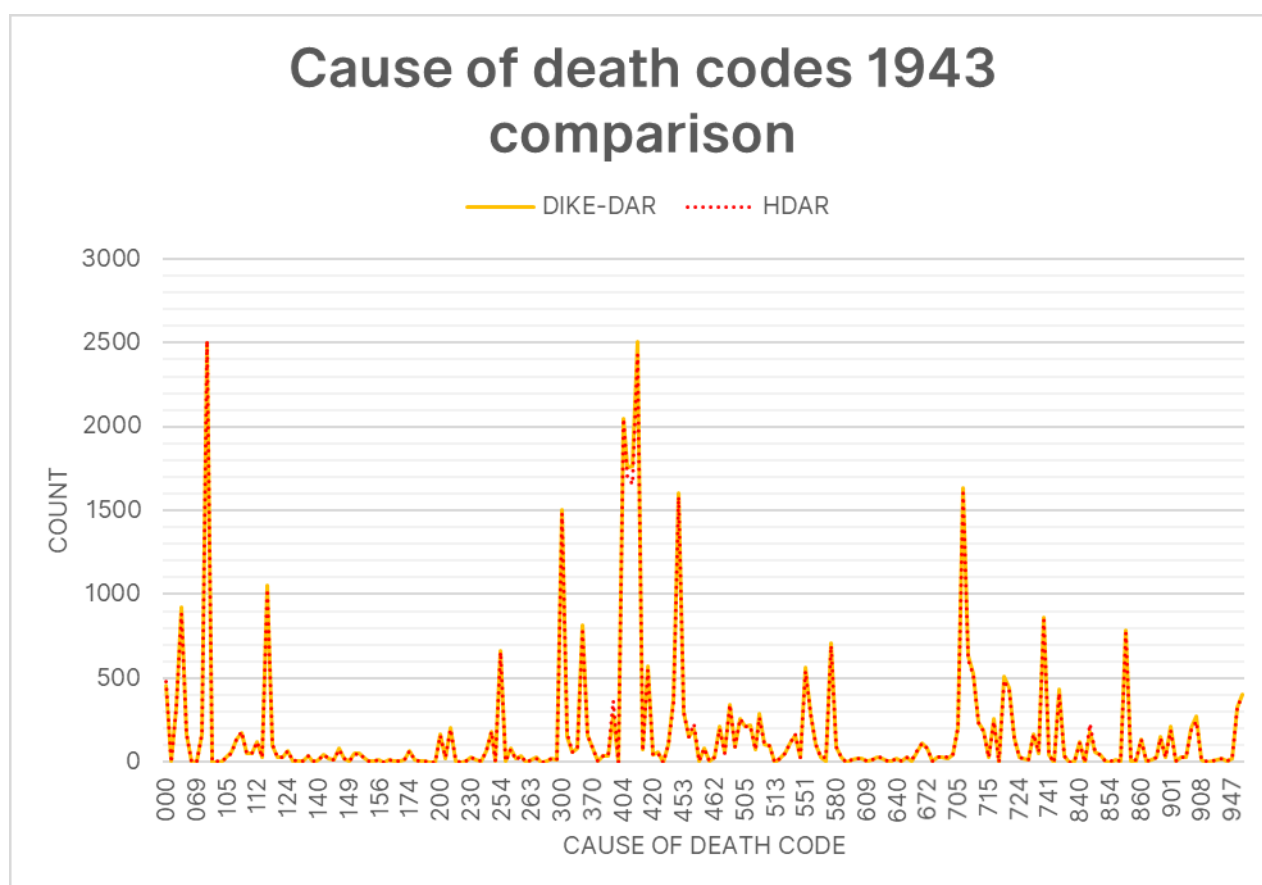


FIGURE 25. COMPARISON DIKE-DAR AND HDAR. CAUSE OF DEATH CODES IN 1943

Finally, the above figure compares the number of people who have died of a specific cause of death code between the two data sets. From this graph we can conclude that – at least for 1943 – the cause of death code is transcribed with very high quality.



5.2.2 Comparison with official statistics

For all the years encompassed in HDAR official death statistics were published by the National Health Authority⁵. These can be used to verify the aggregated number of deaths within the different causes of death. In the section below such comparisons have been made for the years 1941, 1940, 1935 and 1925

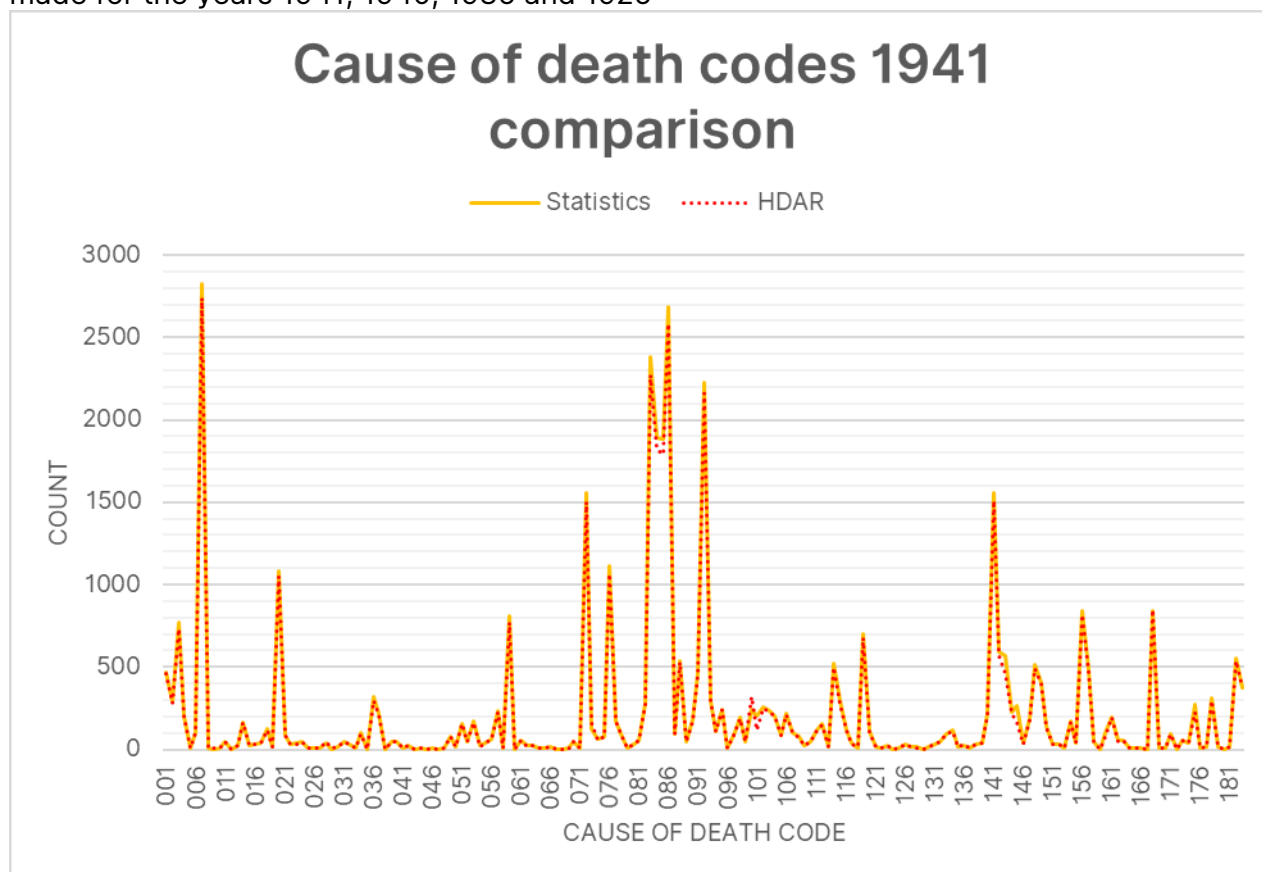


FIGURE 26. COMPARISON OFF. STATS. AND HDAR. CAUSE OF DEATH CODES 1941

The above figure compares the official death statistics for 1940 with HDAR. Some cause of death codes in HDAR (and DIKE-DAR) were not used in the official statistics and for the purpose of this comparison, those have been re-coded as a different code, see Appendix 1.

The average absolute difference between number of deaths of a given cause of death between the two data sets is 9,9. The median absolute difference is 2.

⁵ Dødsårsager i Kongeriget Danmark



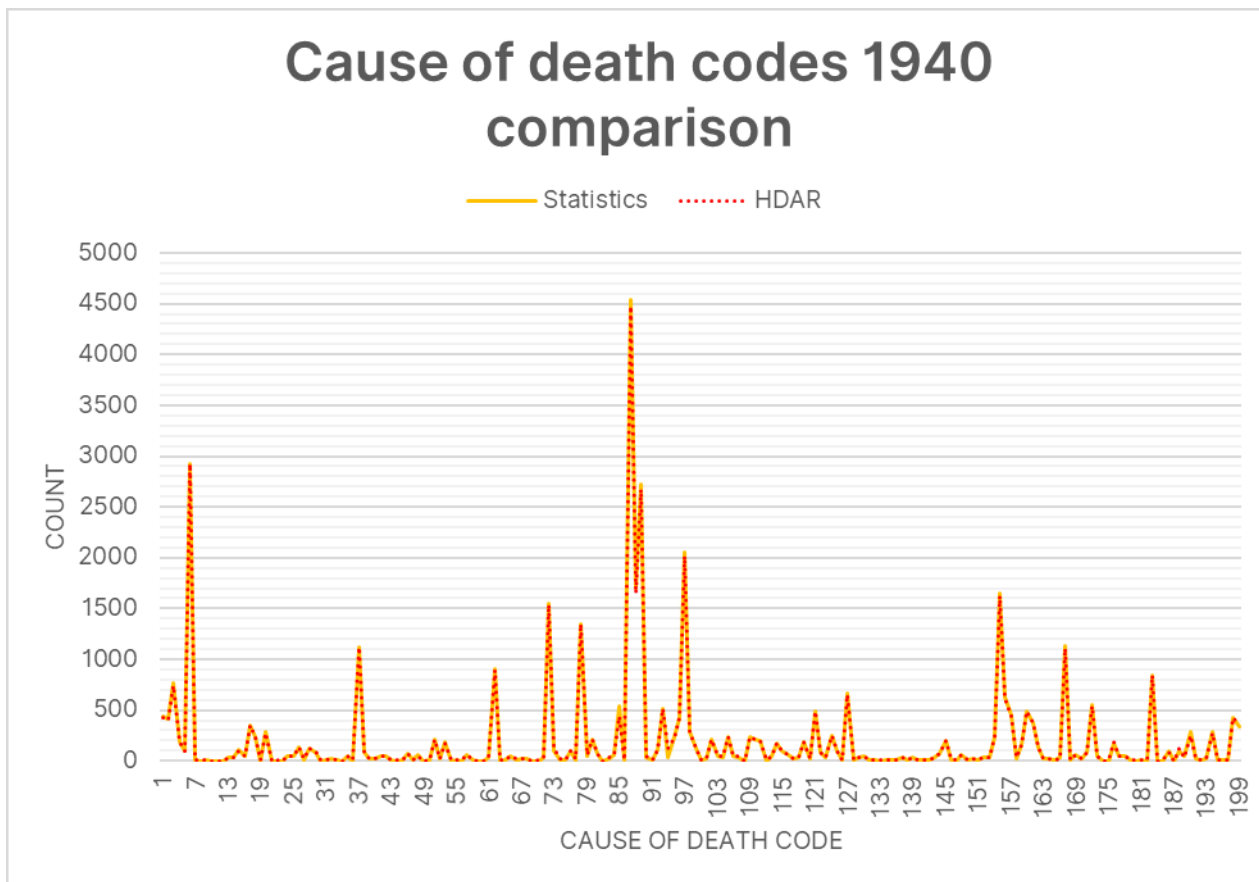


FIGURE 27. COMPARISON OFF. STATS. AND HDAR. CAUSE OF DEATH CODES 1940

The average absolute difference between number of deaths of a given cause of death between the two data sets is 5,5. The median absolute difference is 2.



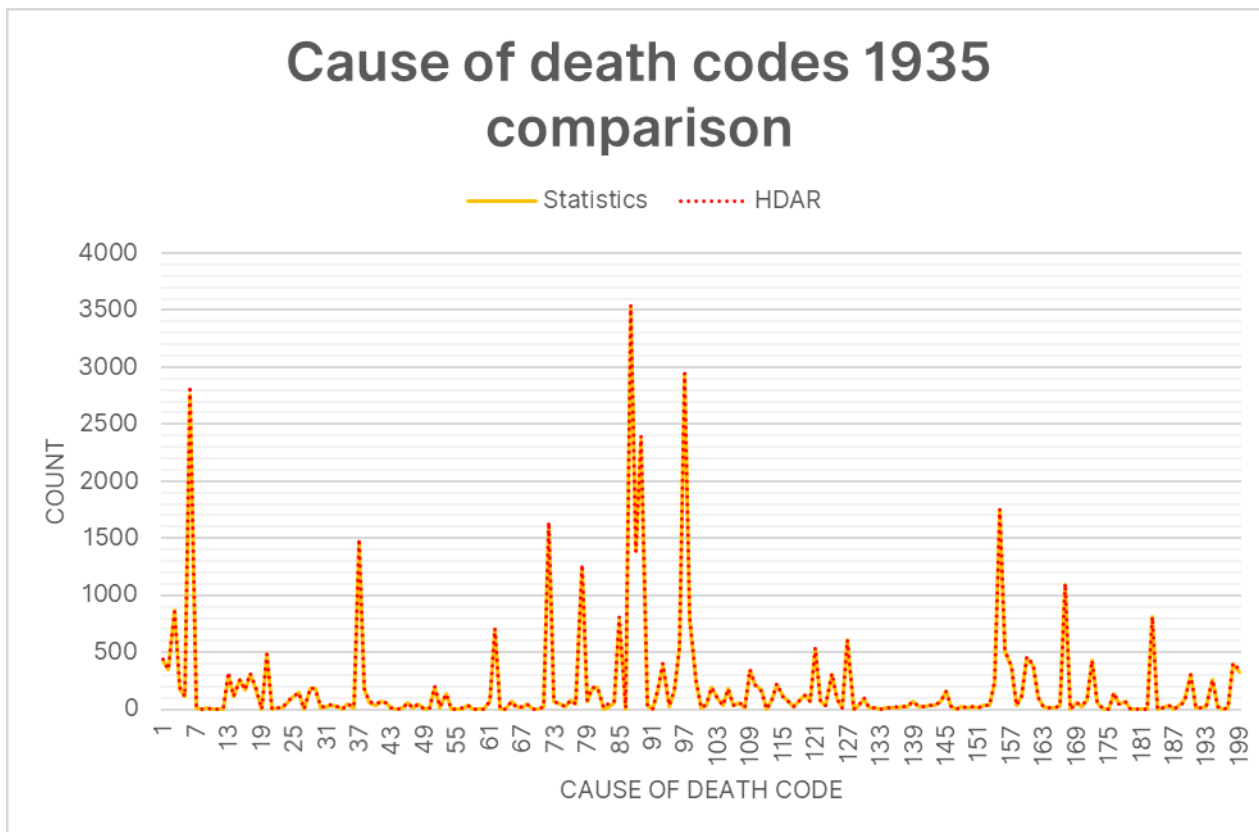


FIGURE 28. COMPARISON OFF. STATS. AND HDAR. CAUSE OF DEATH CODES 1935

The average absolute difference between number of deaths of a given cause of death between the two data sets is 4,9. The median absolute difference is 2.



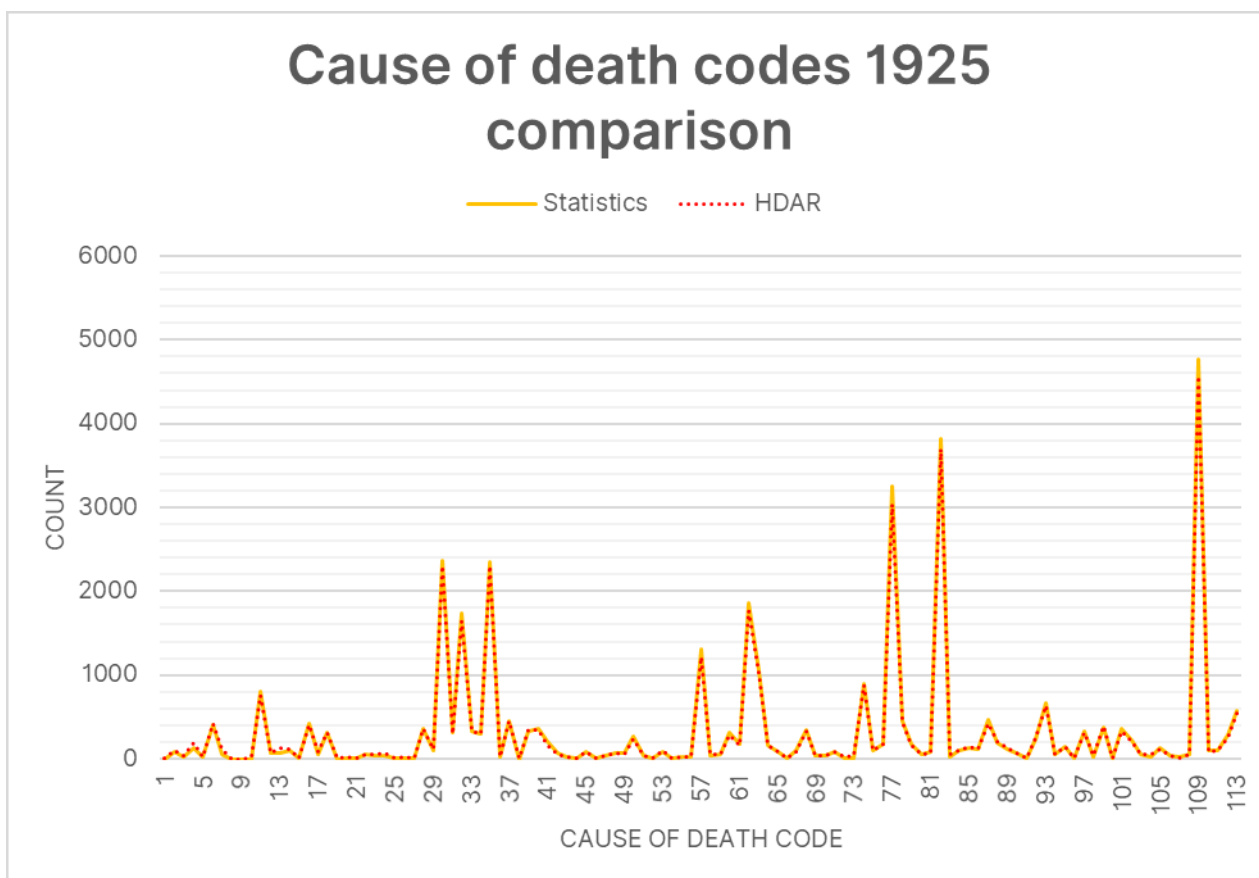


FIGURE 29. COMPARISON OFF. STATS. AND HDAR. CAUSE OF DEATH CODES 1941

The average absolute difference between number of deaths of a given cause of death between the two data sets is 19,7. The median absolute difference is 9. It would seem that the cause of death codes from this early coding period of 1920-1930 has been read more inaccurate than the following notation styles, if the official statistics are believed to be correct. These inaccuracies should be kept in mind if one wishes to analyze or study closer very rare causes of death, where the mean difference of 9 could mean a significant difference in the conclusions.



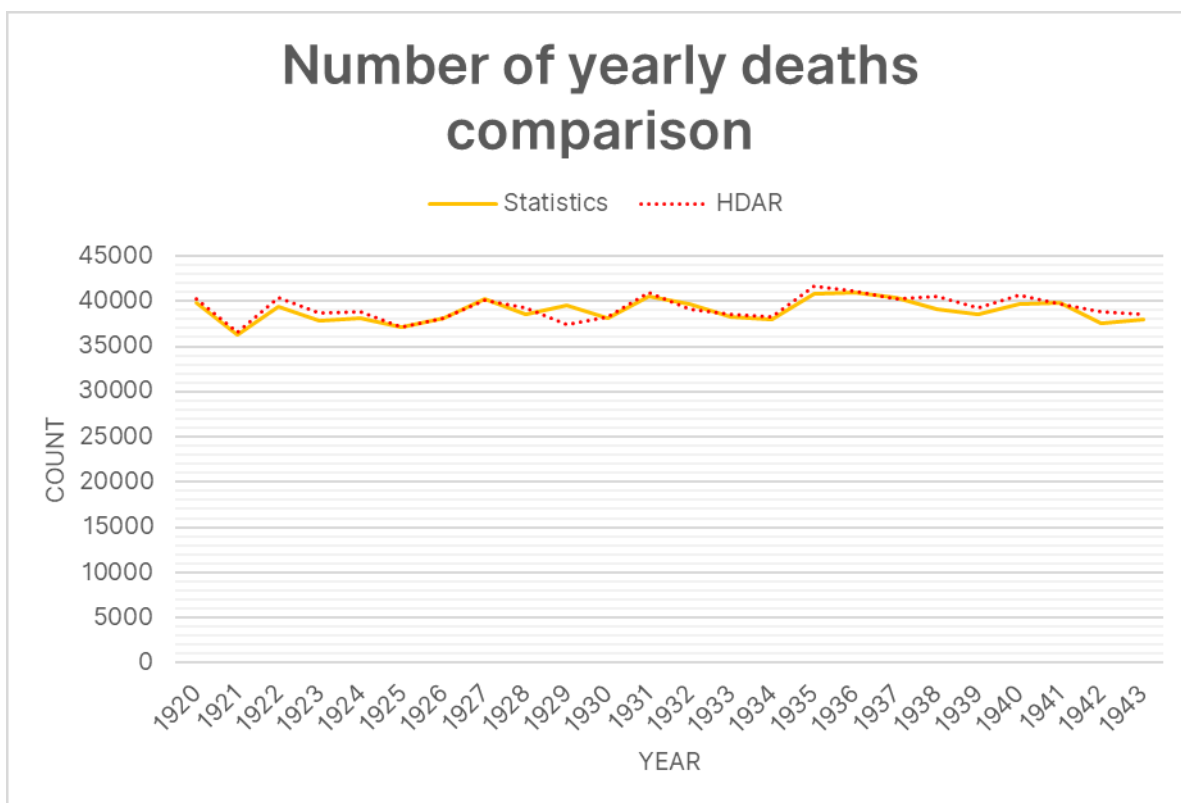


FIGURE 26. COMPARISON OFF. STATS. AND HDAR. YEARLY DEATHS

The average absolute difference between number of deaths of a given cause of death between the two data sets is 586. The median absolute difference is 482. From the graph it is noticable, that in some years HDAR contains more deaths than the official statistics while in others it contains fewer, with a mean difference of about 500 deaths. This difference could be caused by the wrong reading of the death year, using one of the rarer layout formats not included in HDAR (such as deaths in other countries or midwives reports) or duplicate certification which have not been properly identified.

6 Suggested usage and further work

Most of the following points about the usage of HDAR have been mentioned previously in this guide, but are summarized here for convenience.

Through working with the data as well as comparing it with the verification data set, we judge the most reliable transcriptions to be those related to dates, ages, cause of death codes and, to a lesser extent, names. A little less reliable are the free text transcriptions of cause of death, primary illness and manner of death, though we have been unable to quantify just how much, due to the shortcomings of the verification set mentioned in a previous section. The occupation column is considered to be the least reliable. Depending on which columns are relevant for the specific use case, it would be prudent to carry out certain sanity checks, post-processing or evaluation of these less reliable columns.



Most quality assurance efforts have been put into the name and birth date columns in order to make sure that linking the individuals to other data sets is as easy as possible. Linking to another data set could follow the following procedure:

1. Exclude all stillborn, unbaptized and unknowns by filtering **NavnEval** to null
2. Exclude further unknown individuals by filtering **Foedselsdato_date** to not null
3. Match the other data set's date of birth with the value in **Foedselsdato_date**
 - a. Consider starting by filtering out everyone whose birthdate is an approximation based on their age, by filtering **FoedselsdatoEval** to null
 - b. Alternatively, for those with an approximated birthdate as indicated in **FoedselsdatoEval**, consider allowing for matches within a date range equal to $Doedsdato_date - AlderFyldteAar - 364\text{ days}$ to $Doedsdato_date - AlderFyldteAar$
 - c. Consider using the **FoedselsdatoWC75** column instead of **Foedselsdato_date**, in the cases where **FoedselsdatoWC75** has a value.
4. Use the values in **Fornavn**, **Foedenavn** and perhaps **Kaldenavn** for linking names. Consider using **Giftenavn** as well if you are linking to a data set of married women.
 - a. Be advised that **Foedenavn** and **Giftenavn** are constrained to only being a single word. A surname such as "Brunsgaard Trolle" will appear as "Trolle" in **Foedenavn** and "Brunsgaard" as part of the given names in **Fornavn**.
 - b. Allow for a suitable levenshtein distance. The AI-models have a tendency to standardise the spelling of certain interchangeable names, such as the 'K'/'Ch' in "Kristian" or "Christian" which a linking procedure should allow for.
 - c. Consider if it's possible to link on only a certain number of name parts rather than all name parts. Our experience is, that the name transcription models especially starts hallucinating when the deceased had 5+ name parts.
 - d. Consider using the **FuldNavnWC75** column instead of the name part columns in the cases where **FuldNavnWC75** has a value.
5. When matched save the link by the individuals **PersonId**.

The columns **Doedsaarsag** and **Hovedsygdom** could both be verified by applying some standardization algorithms to the free text, making sure that all illnesses are represented in the same way, and flagging if a certain part of the transcription cannot be standardized to any known illnesses. **Doedsmaade** could also be standardized to one of the four expected outcomes: "Naturlig død", "Ulykkestilfælde", "Mord" and "Selvmord". These three columns could be further verified by comparing the transcriptions with the cause of death linked through **FK_DoedsaarsagNomenklatur_DoedsaarsagId**. For most studies on causes of death, we expect it to be easiest to just use the values linked through **FK_DoedsaarsagNomenklatur_DoedsaarsagId**. If one wants to study secondary or tertiary illnesses leading to the death and not just the primary cause of death, it will be necessary to look in the free text transcriptions.

The occupation information in HDAR should be externally verified or otherwise refined before actively using it for research purposes.





Appendix 1 – mapping of 1941 code nomenclature to statistical nomenclature

The table below contains a mapping of inconsistencies between the cause of death codes used on the certificates and the codes used in the statistical publication "Dødsårsager i Kongeriget Danmark" by Sundhedsstyrelsen 1941-1943. Only used for statistical comparison.

DAR code	DAR description	Sundst. code	Sundst. description
145	Lues cerebrospinalis	144	Syphilis acquisita alia
176	Singultus epidemica	175	Alii morbi infectionis
177	Myalgia epidemica	175	Alii morbi infectionis
178	Mononucleosis infectiosa	175	Alii morbi infectionis
179	Morbus Roskildensis	175	Alii morbi infectionis
180	Enteritis typhi murium (Breslau)	175	Alii morbi infectionis
181	Herpes zoster	175	Alii morbi infectionis
182	Psittacosis	175	Alii morbi infectionis
183	Lymfogranulomatosis benigna (Boeck's Sarcoid)	175	Alii morbi infectionis
221	Haemophilia	220	Diatheses haemorrhagicae
222	Purpura	220	Diatheses haemorrhagicae
231	Agranulocytosis	230	Alii morbi syst. Haemopoët. Et sangvinis
233	Morbus Banti	230	Alii morbi syst. Haemopoët. Et sangvinis
273	Beri-Beri	275	Aliae avitaminoses
422	Morbus cordis aortae. (Aortitis luica)	406	Alii morbi cordis
461	Gangraena pulmonum	460	Alii morbi organorum respirationis
462	Abscessus pulmonum	460	Alii morbi organorum respirationis
513	Atrophia hepatis acuta	512	Alii morbi hepatis





581	Salpingo-oophoritis	580	Alii morbi organ. Genital. (non vener.)
641	Embolia in puerperio (non infectiosa)	640	Alii morbi in partu et puerperio
651	Osteomyelitis chronica	650	Osteomyelitis acuta et chronica
741	Sarcoma ossium	740	Cancer alior. Organ. Et localisatio incerta
802	Arsenismus chronicus	801	Aliae intoxicationes chronicae
803	Saturnismus chronicus	801	Aliae intoxicationes chronicae
826	Morphinismus chronicus	801	Aliae intoxicationes chronicae
827	Cocainismus chronicus	801	Aliae intoxicationes chronicae
882	Casus mortiferi in bello orti. (militær incl. brigademedlemmer)	881	Alii casus mortiferi
883	Casus mortiferi in bello orti. (civil)	881	Alii casus mortiferi
906	Suicidium: Kulilte	905	Suicidium: Venificium
907	Suicidium: Sovemidler	905	Suicidium: Venificium
908	Suicidium: Morfin	905	Suicidium: Venificium
909	Suicidium: Denatureret sprit	905	Suicidium: Venificium
910	Suicidium: Arsenik	905	Suicidium: Venificium
911	Suicidium: Fosfor	905	Suicidium: Venificium
912	Suicidium: Stryknin	905	Suicidium: Venificium
913	Suicidium: Nikotin	905	Suicidium: Venificium

The following 5 cause of death codes from 1941- is not mapped to the codes used in "Dødsårsager i Kongeriget Danmark", as it was not obvious where they belonged and there were no certificates using the codes.

DAR code	DAR category	DAR description
103	III. Morbi infectionis	Pestis
104	III. Morbi infectionis	Cholera
113	III. Morbi infectionis	Tularaemi





129	III. Morbi infectionis	Lepra
890	XV. Mors violenta, non naturalis	Executio judiciaria



Appendix 2 – Mapping of extra 1929-1930 codes to 1871 nomenclature

3.316 individuals who died in the period 1929-1930 were assigned cause of death codes that exists neither in the 1871 nor the 1931 nomenclature. In the linking between HDAR and DIM_DoedsaarsagNomenklatur, these codes have been mapped to an approximate 1871 code as defined by the table below. The mapping can contain errors due to misunderstandings.

1929 code	1871 code	1871 description
114	113	Causa mortis vel male vel omnino non indic.
115	007	Febris typhoidea
116	018	Alii morbi epidemici
117	018	Alii morbi epidemici
118	018	Alii morbi epidemici
119	018	Alii morbi epidemici
120	101	Alii morbi organorum interiorum
121	018	Alii morbi epidemici
122	018	Alii morbi epidemici
123	018	Alii morbi epidemici
124	075	Pleuritis, Empyema
125	018	Alii morbi epidemici
126	018	Alii morbi epidemici
127	021	Alia venena animalia
130	035	Cancer in aliis corporis partibus
131	035	Cancer in aliis corporis partibus
132	035	Cancer in aliis corporis partibus
133	035	Cancer in aliis corporis partibus
134	035	Cancer in aliis corporis partibus
135	035	Cancer in aliis corporis partibus
136	035	Cancer in aliis corporis partibus
140	039	Anæmia
141	039	Anæmia





142	039	Anæmia
143	039	Anæmia
144	060	Alia vitia innata
145	101	Alii morbi organorum interiorum
150	087	Enteritis, Colitis, Typhlitis
160	109	Marasmus senilis
161	011	Cholerine & Catarrhus intestinalis acutus
162	101	Alii morbi organorum interiorum
163	101	Alii morbi organorum interiorum
164	101	Alii morbi organorum interiorum
165	101	Alii morbi organorum interiorum
166	068	Ecclampsia
167	101	Alii morbi organorum interiorum
170	102	Phlegmone, Abscessus
171	103	Caries & Necrosis ossium
180	111	In aut brevi post part. mort. (Fb. puerp. excl.)
181	111	In aut brevi post part. mort. (Fb. puerp. excl.)
182	111	In aut brevi post part. mort. (Fb. puerp. excl.)
183	111	In aut brevi post part. mort. (Fb. puerp. excl.)
184	111	In aut brevi post part. mort. (Fb. puerp. excl.)
185	111	In aut brevi post part. mort. (Fb. puerp. excl.)
186	068	Ecclampsia
187	101	Alii morbi organorum interiorum
188	111	In aut brevi post part. mort. (Fb. puerp. excl.)

