Reliability of causes-of-death statistics: the Italian experience from the ICD-10 training course[1]

Francesco Grippo[2], Enrico Grande, Silvia Simeoni, Simona Pennazza, Simona Cinque, Tania Bracci, Luisa Frova

Abstract

Cause of death (CoD) statistics are a major health indicator. One of the most important instrument for improving their reliability is the appropriate use of the ICD-10 as instrument of harmonization and quality. Six research assistants recruited by Istat followed an in-depth coding course and a 8 weeks mentoring period in which they coded 4.050 cases previously coded by experts. The CoD attributed by the trainees was compared with the one attributed by experts. The overall agreement increased during the mentoring reaching the value of 78.4% which is comparable with the literature findings. From the study it emerges the relevance of having accurate and continuous training in order to achieve the best quality for official CoD statistics.

Keywords: ICD-10, mortality coding, cause-of-death statistics, ICD-10 training.

1. Introduction

Cause of death (CoD) statistics are used to monitor the health of populations and are important for health planning and setting priorities for disease prevention. The production of these data is based on harmonized tools and methodologies which allow high comparability of data in time and space. Nevertheless such statistics are exposed to many sources of variability as the completion of the death certificate, the multiple cause coding and the selection of the underlying cause of death. The reliability of CoD coding is an important factor for improving comparability of data at international level and, in order to increase it, many instruments have been developed. Among these the most important are the internationally agreed Classifications including coder's instructions and the automated coding systems (ACS). In this paper, we focus on the variability of CoD coding and how it can be reduced by an appropriate coding training. In particular we describe the experience of the training course provided to recently recruited personnel and the results achieved in terms of coding performance, quality and comparability with official statistics. In this paper, before introducing the methodology and the results, a description of the ICD-10 coding tool and of some international studies for the measure of coding reliability are provided.

[1] The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat.

[2] Istat, *e-mail*: frgrippo@istat.

1.1 The ICD-10: a tool for classifying mortality and morbidity data

The International Statistical Classification of Diseases and Related Health Problems - 10th Revision (ICD-10) is the standard diagnostic tool for epidemiology, health management and clinical purposes (WHO, http://www.who.int/classifications/icd/en/). It belongs to the International Classification Family, edited by the World Health Organization (WHO) and used for the description of health related topics (WHO, 2009). ICD-10 was adopted by the World Health Assembly in 1990. The ICD was historically developed for coding causes of death but, since 1948 (the sixth revision) it was also used for coding morbidity data. The ICD provides a system of organized categories representing different morbid entities. These categories are identified by an alphanumeric code which allows the standardization of terminology, the organized capture, memorization and the systematic analysis of morbidity and mortality data.

The ICD-10 is a dynamic tool, with an annual updating process that allows to keep up with the continuous advances in medical sciences and to ensure the best use of it. The ICD-10 is published in three different volumes: volume 1 contains the organized list of categories and other tools such as lists of tabulation; volume 2 encloses definitions and application rules; volume 3 represents the alphabetical index with the medical terminology and the related ICD-10 code. In Italy, the ICD-10-version 2009 is currently adopted and it is available for online browsing at Istat website (Istat, 2014).

1.2 Cause of death coding: complexity of rules

The starting point of CoD statistics is the death certificate where the causes of death are notified. In Italy, the certificate is filled out by a medical doctor who generally knows the medical history of the deceased (family doctor or attending physician at hospital), but can also be filled by necropsy physician when attending is not available. The Italian death certificate follows the structure of the international one, provided in the ICD-10, which consists in two parts: in the first part, the physician should report the complete sequence of conditions directly leading to death; in the second part, the other conditions contributing to death.

In the death certificate many conditions can be reported, but to allow comparison of cause of death statistics within and between countries, only one cause is selected: the underlying cause of death (UCD). The concept of UCD was introduced with the sixth revision of the ICD and corresponds to "(a) the disease or injury which initiated the train of morbid events leading directly to death, or (b) the circumstances of the accident or violence which produced the fatal injury".

The Classification provides a set of mortality coding rules that allows the standardization of the underlying cause of death selection and the choice of ICD code that better fulfills the definition of underlying cause, using all the information reported in the death certificate.

The ICD rules can be divided into two groups: the selection and the modification rules. The selection rules are *General principle*, *Rule 1*, *Rule 2* and *Rule 3* and they are used to identify the originating antecedent cause that initiates the sequence of conditions leading to deaths. The modification rules, rules A trough F, are used to modify the first selected condition in order to get a more relevant and informative code. The rules are described in the ICD-10 volume 2, where examples, applicability and comments for the application are also provided (WHO, 2009).

In practice, the application of rules presents various problems. When applying General Principle, Rule 1 and Rule 2, the coder should analyze the sequence of conditions reported by the certifier and decide if each condition is in a correct causal relationship with the others. On the other hand, when applying Rule 3, the coder should decide if the condition temporarily selected as underlying cause could be considered an obvious consequence of others conditions reported. Moreover modification rules are provided for giving a more detailed, specific and relevant information. In this case, the coder should know which condition is the most relevant in order to select it.

The ICD volume 2 provides guidelines for the application of coding rules. Despite this, some instructions could give rice to personal interpretation. An international tool for limiting this problem is represented by a set of decision tables used also by the automated coding systems. These tables were developed by the US National Institute for Health Statistics (NCHS, 2009) as part of the ACME software (CDC, 2007), a tool for the automated selection of the underlying cause of death. Successively, also in Europe, the tables have been maintained as part of Iris software (Iris, 2014), a new coding tool which integrates all phases of automated coding (text recognition, coding of each single diagnostic terms and the selection of the underlying cause) and allows an interactive handling of the rejects.

The prerequisite for the correct application of decision tables is the coding of each condition present in the death certificates. This is a complex task because different ICD-10 codes can be attributed to the same condition depending on different variables, such as the age and gender of the deceased, the duration of each condition (interval between onset of diseases and death) and the presence of other conditions on the certificate.

The complexity of coding is increased by the fact that, besides the described coding rules, other special instructions are provided for specific cases such as: perinatal and infant mortality, congenital conditions, external causes, complications of surgery.

1.3 Problems and measures of cause-of-death coding reliability: the international experience

The complexity of coding rules requires that coders should be deeply trained in the use of ICD-10. Intensive training coding course are necessary to increase the competence of coders and consequently the accuracy of CoD coding. To evaluate the effectiveness of training course and generally the reliability of CoD coding, different methods and statistical indicators are proposed by many authors. In this paragraph the strengths and weaknesses of different methods are reported.

The coding process is expected to be independent of the coding person, coding time and space. Nevertheless, despite the detailed and specific rules provided in the ICD, even coding experts show different opinion in selecting the underlying cause of death (Buchalla C. et al. 2013). This fact leads some authors to define the use of ICD for coding a “matter of chance” (Stausberg J. et al., 2008). The same Authors refer that some coding errors are due to the intrinsic limitations of ICD-10 which actually includes some ambiguities and inconsistencies.

Errors in ICD-10 coding can derive from different sources such as: (1) the incorrect and/or incomplete reporting of the causes on the death certificate by the physician, (2) the complexity of medical nomenclature and national language, (3) the interpretation of coding or selection rules, (4) individual deliberation of coders.

Actually the information available on the death certificate and how the physicians report it, is crucial for a proper coding. According to our experience, certifiers often report more than one underlying cause, despite the recommendation; moreover the reported condition often corresponds to the immediate cause or complications of the actual UCD (Grippo F et al. 2013). All these factors make arduous to properly apply the ICD-10 coding rules and to avoid the personal interpretation.

An important source of UCD variability is related to the complexity of medical nomenclature and its interpretation by coders, who normally are not medical doctors. Moreover, the use of medical terms in national language can be different, leading to national or even regional differences.

The interpretation of ICD rules and guidelines is not unequivocal ant it leaves room for individual choices (Stausberg J. et al. 2008). For most cases, the above discussed ACME decision tables provide the correct way for rules application. Nevertheless, tables do not cover all textual instruction. This is especially the case of special instructions applied when the death certificate reports complications of surgical intervention. For these instructions the ACME tables cannot be used as a reference guide. Errors that can affect the selection of the underlying cause of death can be labeled as miscoding and misspecification (O'Malley K.J. et al., 2005). Miscoding occurs when the underlying cause code is misaligned with the evidence found in the death certificate. Misspecification includes assignment of generic codes when information exists for assigning more specific codes.

The reliability of cause of death coding can be evaluated by different methods divided mainly into two groups: the ones that use gold standard (GS) and the others that don't use gold standard (NGS). In the first group, the underlying cause attributed by each coder is compared with GS, generally the UCD coded by a reviewer (Lu T.H. et al., 2000). In the second group, many coders code the same certificates and the UCDs are compared with each other (Harteloh P. et al., 2010).

The indicators used are: the percentage of agreement P (i.e. the percentage of death certificates for which all coders (or between each coder and the reviewer) give the same UCD) and the K statistic (Cohen J., 1960) generally thought to be a more robust measure than the simple percent agreement calculation since it takes into account the agreement occurring by chance. Indicating with P the relative observed agreement among raters, and with Pe the hypothetical probability of chance agreement, the K statistics is calculated as $P\text{-}Pe/1\text{-}Pe$. When the raters are in full agreement, then K=1. If there is complete disagreement, then K=-1; if there is independence among the raters K=0. Besides the K statistic, other indicators are used as false positive and false negative rates.

In Lu's article (2000) the underlying cause attributed by each coder is compared with GS: 5,621 death certificates were re-coded by an expert reviewer. The UCD selected by the expert was treated as GS and used to calculate the agreement rate and the K value. The overall agreement rates between the reviewer and coders according to the 3 digit and 2 digit categories of ICD-9 were 80.9% and 83.9%. The percentage of agreement decreases with the number of conditions per certificate and the age of deceased but not significant differences were observed by sex. Higher agreement was found for malignant neoplasms (K=0.94) and injuries and poisoning (K=0.97), but there was poor agreement for nephrotic diseases (K=0.74), hypertension-related diseases (K=0.74), and cerebral infarction (K=0.77).

In Harteloh's article (2010), the authors study the reliability of cause of death statistics in the Netherlands, calculating the percentage of agreement among coders (method NGS).

The percentage of agreement is measured as the percentage of death certificates for which all coders (four) give the same UCD. They calculated the inter-coder agreement, by comparing the UCD of each death certificate attributed by different coders and the intra-coder agreement, by comparing the UCD attributed by the same coder in different periods. 10,833 death certificates, already coded, were manually re-coded by four coders. The intra-coder agreement was 88–90% at a 4 or 3 digit level and 95–96% at chapter level. It was the same in magnitude as the inter-coder agreement for pairs of coders (87% at a 4 digit, 89% at a 3 digit and 94% at chapter level) and the authors concluded that "the coding process in itself has limited reproducibility and is not bound by individual preferences of coders". The agreement of coding process was associated with the level of detail of the ICD-10 code (chapter, 3 digit, 4 digit), the age of the deceased, the number of coders and the number of diseases reported on the death certificate. The reliability of cause-of-death statistics turned out to be high (90%) for major causes of death such as cancers and acute myocardial infarction. For chronic diseases, such as diabetes and renal insufficiency, reliability was low (70%). These conditions are associated to higher number of diseases per certificate and older age of deceased and this factors can contribute to a major variability.

It is difficult to compare different studies of reliability because of the variety of protocols and measures (different number of coders, different statistical indicators, etc.). Nevertheless, it is possible to draw some general conclusions: what emerges is that the coding reliability is lower when certificates report chronic diseases such as diabetes, hypertension related diseases, chronic liver diseases, etc., because they are long term diseases associated with old age of deceased and they are part of very complex morbid patterns. Value of coding reliability indicators also decreases with the number of codes per deceased (that increases with the age). It is necessary to understand better the coding process weaknesses and to increase the reliability of coding by an adequate training course of the coders and clearer instructions provided by the ICD, especially for some cases.

1.4 The cause of death coding process adopted in Italian National Institute of Statistics (Istat)

In Italy Istat is in charge of the cause-of-death coding. Each year, around 600 thousands death certificates are collected by Istat and electronically recorded. The certificates are processed through an automated coding system (ACS) and the rejected ones are manually coded by expert coders.

ACS process can be divided into three different steps.

Step1. The death certificates are analyzed by ACTR, a software for text recognition (Wenzowski, 1988) that transforms each recognized entry (diagnostic term) into a standardized code (Entity reference number - *ERN*) .

Step 2. The second software, MICAR (Mortality Medical Indexing, Classification, and Retrieval) converts *ERNs* in the correct ICD-10 code.

Step3. The third software, ACME, automatically applies the international rules of the ICD-10 and selects the underlying cause of death.

Rejects can be produced in each step of this process. When a certificate is rejected, manual coding is necessary. If the reject occurs for failure in step 1 or 2 the manual coder can either correct the ICD codes attributed to each rejected diagnostic term (multiple cause coding) and then submit the certificate to ACME, or can manually select the UC. Rejected in the step 3 are only handled in this second modality.

About 80% of death certificates is fully automatically coded, the remaining 20% are rejected and manually coded. The rejects are more complex than the other certificates. In fact, certificates with many diagnostic terms have more probability to be rejected and the automated coding cannot handle some complex cases such as surgery deaths, external causes, deaths mentioning drug therapies.

2 Methods

2.1 The training

Six research assistants were recruited by the Italian National Institute of Statistics for the cause of death unit. They have a university education in statistics, mathematics or biological sciences.

They followed a coding training course divided into ICD-10 lectures structured in three modules:

Module 1) Cause of death statistics: the data workflow and the use of cause of death data, one day duration: six hours;

Module 2) Coding and selecting causes of death using ICD-10 version 2009; 11 days: 71.5 hours;

a. part 1: selection and modification rule
b. training on the job on selected cases (one week),
c. special cases (external causes, complication of surgery and medical therapy, rheumatic heart diseases, infant deaths, drug poisoning, interpretation of death certificate)

Module 3) Software tools for coding: two days: 10 hours.

The lectures period lasted from January to March 2013. The reference manual for the course was prepared by Istat (2010) as an extensive integration of ICD-10 volume 2, based on NCHS manual 2a (2007) and referring to ICD updates until 2009 (Istat, 2010). The WHO training tool (2012) was also consulted.

The teachers of the course were the senior coders of the Istat cause of death unit with a long experience in ICD-10 mortality coding.

After the training course, the six research assistants (i.e. trainees) underwent also a period of mentoring lasting 17 weeks (from March to June 2013). During this period, each trainee coded real cases and was supervised by senior coders. Periodic meetings were organized in which coding doubts were clarified and some cases were revised. Coding results were evaluated and monitored to individuate possible errors of application or misunderstanding of international coding rules.

2.2 Evaluation of learning process

During the mentoring period, each trainee coded the same set of 4,050 death certificates rejected by the automated coding system. These certificates, referring to deaths occurred in Italy during the month of December 2010, had been previously coded by senior coders of Istat during the routine data processing. As discussed in the introduction, these certificates can generally be considered more complex than those fully automatically coded.

The UCD attributed by senior coders was taken as gold standard. Certificates were the same for all the trainees and, at the end, a total of 24,300 deaths certificates were available

for the analysis. The coding was computer assisted and requested the completion of multiple causes (MC) i.e. the complete coding of each condition reported on the death certificate. For certificates with complete MC, ACME software was used to select the UCD. Manual selection was performed on certificates with incomplete MC, certificates containing complications of surgery or external causes.

As previously reported in literature, many indicators of coding reliability have been used for different settings and purposes. For our objective the best indicator had to provide a direct and summary measure of misalignment between the coding performed by the trainees and the standard coding practices adopted by the Istat senior coders (gold standard). We did not use the *K* as it is designed to measure the degree to which the different coding choices agree with each other (precision) rather than the accuracy of the choice (closeness to the gold standard) (Viera A.J., 2005; Kwiecien R., 2011). Moreover, as reported in the literature, the definition of chance agreement is highly controversial (Brennan, R. L., 1981) and often not applicable in practice. In our case the probability of attribution of the same code due to chance is very low (close to 0, as the number of ICD-10 attributable codes is about 10,000). This makes the values of the raw proportion of agreement *P* very close to the *K* values, so we chose the first indicator as it enables more immediate interpretation of the results.

Therefore we used the indicator P_i defined, for each trainee *i*, as the proportion of certificates for which there was an agreement on the UCD with the senior coders.

The basic formula for the agreement P_i was:

$$P_i = \frac{n_i}{N_i}$$

where n_i was the number of certificates coded by the trainee *i* with UCD that agreed with the one attributed by the senior coder, and N_i was the total number of certificates coded by the trainee *i*.

The 95% confidence intervals for the agreement P_i was calculated as follows:

$$P_i \pm 1.96 \times \sqrt{\frac{P_i \times (1-P_i)}{N_i}}$$

The overall agreement *P* for the all six trainees combined was

$$P = \frac{\sum_i n_i}{\sum_i N_i} \qquad \text{for } i=1,\ldots,6.$$

The overall agreement was calculated at different level of detail of the ICD-10 classification: at 4 digit level; at 3 digit level and at group level.

A time-trend evaluation of the agreement was performed by calculating the indicator weekly.

The cause of death agreement was calculated by grouping the certificates according to the UCD coded by senior coders. Conforming to this approach, the proportion described above was calculated for a specific set of certificates with cause of death *c* selected as UCD (N^c):

$$P_i^c = \frac{n_i^c}{N^c}$$

In our study *c* indicated a broad category of causes of death, such as ICD-10 chapters, or specific coding topic, i.e. sequelae or rheumatic heart diseases.

To make the interpretation of results easier, the agreement was expressed as percentage.

The daily number average of coded certificates was used to monitor the increase of work rhythm during the mentoring period.

An additional analysis was carried out in order to investigate the agreement of certificates containing medical procedures. The total set of death certificates coded during the mentoring was divided into two groups: certificates containing mention of surgery and other medical procedures (781 deaths) and the certificates not containing it.

3 Results

The average number of certificates coded per day by each trainee increases during the mentoring period from 18 to 96 (table 1). Actually, this is an expected result as the ability of coding increases with the practice. Of all the 4,050 death certificates coded by each trainee, the overall agreement with the senior coders is 78.0% at 4 digit level and 82.3% at 3 digit. Both these values increment significantly over time: at 4 digit level it passes from 70.8% to more than 78.4%. The maximum value of agreement is reached in about 7 weeks (80.1%).

Table 1 – Overall agreement by mentoring week, between trainees and senior coders, at 4 and 3 digit level

Week	Number of certificates	Person – day	Average certificates coded by each trainee per day	Overall agreement P at 4 digit					Overall agreement P at 3 digit		
		(N)	(N)	%	IC95%		maximum	minimum	%	IC95%	
					inf	sup				inf	sup
1-2	1.177	65	18,0	70,8	68,2	73,4	64,9	78,7	77	74,1	79,8
3-4	1.299	56	23,0	75,1	72,7	77,5	69,1	79,8	79,9	77,4	82,4
5-6	1.542	43	36,0	77,2	75,1	79,3	66,9	88,9	82,2	79,9	84,4
7-8	2.894	57	51,0	80,1	78,6	81,6	76,2	88,2	84,2	82,7	85,8
9-10	3.143	37	85,0	78,6	77,2	80,0	75,3	84,0	83,3	81,8	84,8
11-12	4.595	55	84,0	77,6	76,4	78,8	71,5	82,5	81,7	80,4	83,0
13-14	3.308	51	65,0	79,1	77,7	80,5	74,5	83,5	84,0	82,5	85,5
15-17	6.342	66	96,1	78,4	77,4	79,4	74,2	85,4	82,1	81,2	83,1
Total	24.300	430	57,0	78,0	77,5	78,5	64,9	88,9	82,3	81,8	82,9

Thereafter the level of agreement is quite stable until the end of the mentoring period. At the same time, the variability of the agreement by trainee results to decrease: the maximum and minimum values of the agreement are observed to converge on high agreement levels (table1).

In table 2 the agreement for each trainee is presented: it has a range of variation of about 4 (from 75.8% to 79.7%). In the first four weeks this range is greater (about 6, ranging from 69.7% to 75.9%) and reaches the value of 5 in the last four weeks. The trainees which started with a lower rate of agreement compared to the others, show a greater improvement (difference between first and last four weeks). The agreement variability among the trainees decreases from 6 to 4 percent points.

Table 2 – Agreement by trainee during the first and last four mentoring weeks, at 4 digit

Trainee	Overall agreement			First four weeks			Last four weeks			Difference between first and last four weeks
	%	IC95%		%	IC95%		%	IC95%		
		inf	sup		inf	sup		inf	sup	
1	79.7	78.4	80.9	74.1	69.6	78.5	80.8	78.9	82.7	6.8
2	75.8	74.5	77.1	69.7	64.9	74.5	76.7	74.6	78.9	7.0
3	78.4	77.2	79.7	72.6	68.0	77.1	78.7	76.7	80.8	6.2
4	75.9	74.6	77.2	72.9	68.8	77.0	76.7	74.8	78.6	3.8
5	78.2	77.0	79.5	72.4	68.2	76.6	79.5	77.5	81.5	7.1
6	79.7	78.5	80.9	75.9	72.1	79.6	79.8	77.7	81.8	3.9
Total	78.0	77.5	78.5	73.1	71.3	74.8	78.7	77.8	79.,5	5.6

Another interesting result is the agreement at 4 digit by groups of UCD which has significantly different values (table 3): it varies from a value of 36.8% for medical procedures and therapies to 87.9% for congenital malformations and chromosomal abnormalities. Besides this latter group, a significantly higher value of the agreement is observed for type II diabetes mellitus, dementia and endocrine, nutritional and metabolic diseases (except diabetes) for which the value of this indicator is higher than 85.5%. For other groups of causes such as symptoms and signs; malignant neoplasm; chronic liver diseases; diseases of the nervous system and circulatory diseases (except rheumatic) the value ranges between 79.6% and 84.3%.

On the other hand, alongside mentioned medical procedures and therapies, the group of mental and behavioral disorders (excluded dementia); sequelae codes; infectious diseases (other); diseases of blood and blood forming organs; transport accidents and diseases of the digestive system (other) show low values of the agreement between 57.9% and 70.0%. For rheumatic heart diseases; other valvular diseases; diseases of the genitourinary system and external causes the agreement has a value between 70.4% and 76.1%.

The overall agreement at group level has a value of 90% (IC95% 89.6-90.3). The analysis of this indicator by cause of death confirms the results discussed at 4 digit level. Nevertheless there is a difference for external causes and transport accidents: the agreement at 4 digit shows lower values compared to the average, while the agreement at group level is significantly higher. This finding indicates that for these causes of death there are difficulties in attributing the appropriate 4 digit, but the cause of death remains classified in the same group.

A more detailed analysis of the agreement between trainees and senior coders can be drawn from table 4, where the UCD attributed by trainees, at group level, is cross-tabulated against the gold standard UCD. On the diagonal of the table, there is the percentage of certificates that are coded in the same group by trainees and senior coders (agreement at group level already presented in table 3). Figures outside the diagonal represent the percent of certificates which are attributed to another cause by the trainees compared to the gold standard. The additional information provided by this table is the possibility to evaluate the direction of the different classification between trainees and senior coders, i.e. the percent of cases allocated to a different group by the trainees. For example, from the table, it is evident the misclassification between viral hepatitis and chronic liver diseases; non-malignant neoplasm and malignant neoplasm; rheumatic heart diseases and other valvular diseases; medical procedures and sequelae.

The inter-coder variability of the agreement has a different behavior according the UCD, as shown in figure 1, where the agreement by trainee (continuos line) is compared with average one (dotted line). Distinct scenarios are presented. In correspondence of causes of death with high rate of agreement, low variability is

observed for neoplasms, while there is a certain degree of variability for congenital malformations (with two trainees out of six reaching the 100% level of agreement). At intermediate level of agreement, it is possible to observe low variability pattern (respiratory diseases) or high degree of variability (viral hepatitis). At last, the results for rheumatic heart disease and especially for sequelae represent two examples of how the inter-coder variability varies in correspondence of lower level of agreement.

Table 3 – Overall agreement at 4 digit level and at group level by underlying cause of death groups

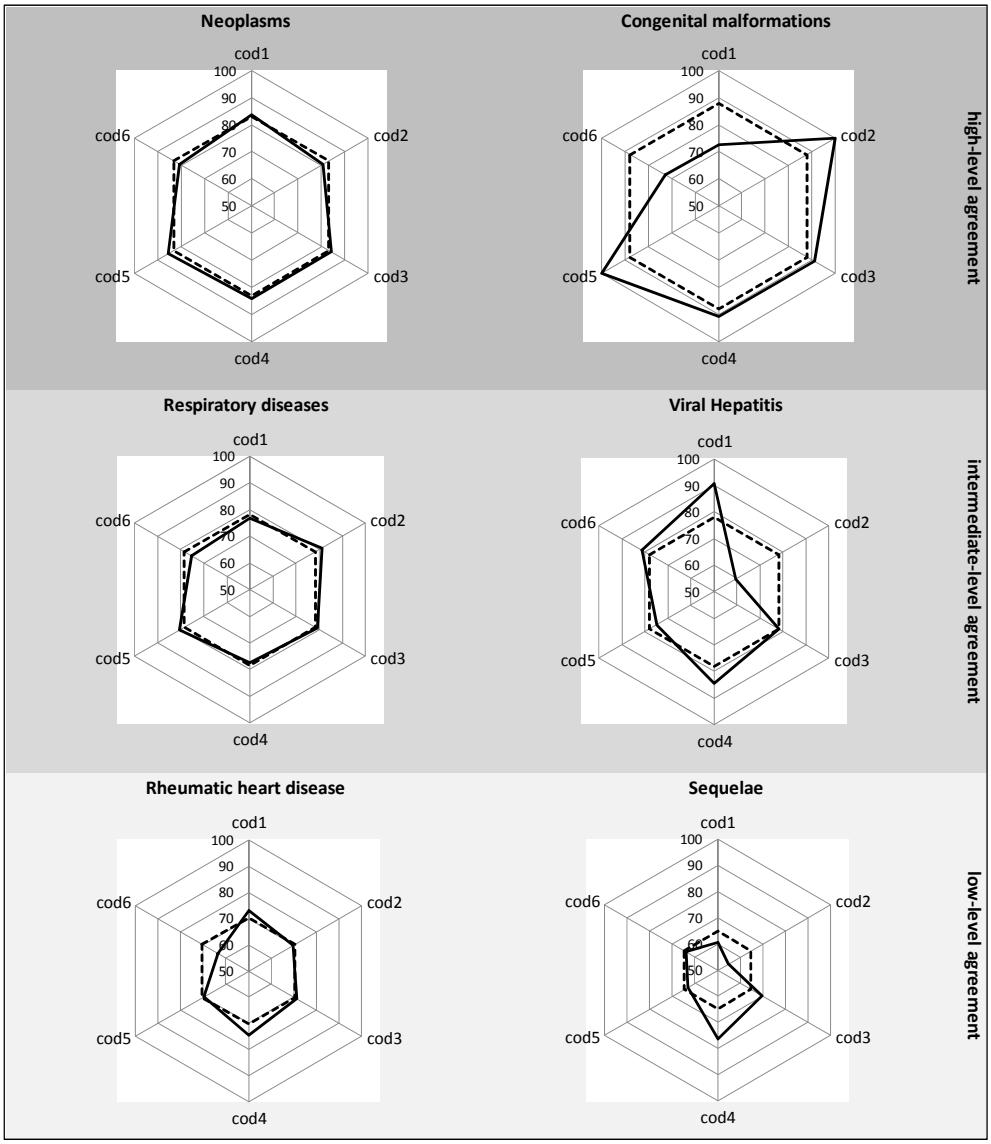
Cause of death groups	Number of certificates	Agreement at 4 digit			Agreement at group level		
			IC95%			IC95%	
		%	Inf	Sup	%	Inf	Sup
Viral hepatitis	192	78.1	72.3	83.9	83.9	78.7	89.1
Infectious diseases (other)	306	67.6	62.4	72.8	80.7	76.3	85.1
Malignant neoplasm	5,880	83.4	82.4	84.4	95.5	95.0	96.0
Other neoplasms	438	79.7	75.9	83.5	87.0	83.9	90.1
Diseases of blood and blood forming organs	162	67.9	60.7	75.1	73.5	66.7	80.3
Type II diabetes	138	86.2	80.4	92.0	93.5	89.4	97.6
Diabetes (Other)	480	80.0	76.4	83.6	88.5	85.6	91.4
Endocrine, nutritional and metabolic diseases (other)	372	85.5	81.9	89.1	92.5	89.8	95.2
Dementia	342	86.0	82.3	89.7	87.4	83.9	90.9
Mental and behavioural disorders (other)	114	57.9	48.8	67.0	72.8	64.6	81.0
Diseases of the nervous system	870	79.9	77.2	82.6	86.8	84.6	89.0
Rheumatic heart diseases	378	70.4	65.8	75.0	87.0	83.6	90.4
Other valvular diseases	444	71,2	67.0	75.4	77.7	73.8	81.6
Circulatory diseases (other)	6.45	79.6	78.6	80.6	91.1	90.4	91.8
Respiratory diseases (other)	1,140	78.3	75.9	80.7	88.7	86.9	90.5
Chronic liver diseases	306	82.4	78.1	86.7	85.9	82.0	89.8
Diseases of the digestive system (other)	924	69.9	66.9	72.9	81.0	78.5	83.5
Diseases of the genitourinary system	300	72.0	66.9	77.1	83.7	79.5	87.9
Congenital malformations, deformations and chromosomal abnormalities	66	87.9	80.0	95.8	92.4	86.0	98.8
Symptoms, signs and abnormal clinical and laboratory finding	108	84.3	77.4	91.2	89.8	84.1	95.5
Transport accidents	1014	69.6	66.8	72.4	96.7	95.6	97.8
Medical procedures and therapies	144	36.8	28.9	44.7	58.3	50.2	66.4
External causes (other)	2,670	76.1	74.5	77.7	93.8	92.9	94.7
Sequelae	732	64.8	61.3	68.3	70.2	66.9	73.5
Other	330	63.9	58.7	69.1	74.8	70.1	79.5
Total	24,300	78.0	77.5	78.5	90.0	89.6	90.3

Table 4 –Cross tabulation of underlying cause of death selected by senior coders and trainees*

Cause of death attributed by trainees	Cause of death attributed bySenior coders (gold standard)																								
	Viral hepatitis	Infectious diseases (other)	Malignant neoplasm	Other neoplasms	Diseases of blood and blood forming organs	Type II diabetes	Diabetes (Other)	Endocrine, nutritional and metabolic diseases (other)	Dementia	Mental and behavioural disorders (other)	Diseases of the nervous system	Rheumatic heart diseases	Other valvular diseases	Circulatory diseases (other)	Respiratory diseases (altre)	Chronic liver diseases	Diseases of the digestive system (other)	Diseases of the genitourinary system	Congenital malformations, deformations and chromosomal abnormalities	Symptoms, signs and abnormal clinical and laboratory finding	Trasport accidents	medical procedures and therapies	External causes (other)	sequele	Other
Viral hepatitis	83,9		0,0		3,7					1,8						8,2									
Infectious diseases (other)	0,5	80,7	0,0	0,5			0,4				0,3			0,2	0,1		4,0					0,7		0,8	2,7
Malignant neoplasm		2,0	95,5	5,3	3,1		0,4	0,3	0,3		1,7	0,3	0,2	0,2	1,2	2,6	0,9	0,3				3,5	0,1	0,4	
Other neoplasms			1,3	87,0	0,6			0,3			0,2		0,2	0,1	0,3		0,2	1,7					0,0		
Diseases of blood and blood forming organs			0,2	0,7	73,5					0,9	0,3			0,1								3,5	0,1		
Type II diabetes						93,5			0,3		0,1			0,0			0,3								
Diabetes (Other)	1,6						88,5		0,6		0,5			0,4	0,1								0,2	0,5	
Endocrine, nutritional and metabolic diseases (other)	1,0		0,2	0,2	1,9		0,6	92,5		2,6	0,1			0,1	0,2	0,7				0,9				0,4	1,2
Dementia		0,3							87,4					0,1	0,4		0,8								
Mental and behavioural disorders (other)				0,2			0,2		0,6	72,8	0,1			0,0		0,3	0,4						0,2	0,3	
Diseases of the nervous system		2,0	0,0	0,5	1,2		0,6	1,3	2,3	2,6	86,8	0,3		0,8	0,7		0,1					3,5	0,8	0,8	3,0
Rheumatic heart diseases	0,5	0,7			3,7			0,3				87,0	10,8	0,6	0,1		0,1					1,4			0,9
Other valvular diseases			0,0		1,9		0,4					4,8	77,7	0,5			0,1							0,1	
Circulatory diseases (other)	2,1	1,6	0,6	2,1	1,9	2,2	4,0	1,9	4,1	2,6	2,9	5,6	7,2	91,1	3,4		4,1	3,0	3,0	1,9	0,2	9,7	0,7	10,1	5,2
Respiratory diseases (altre)		3,3	0,4	0,2							0,6		0,9	0,5	88,7				1,5			0,7	0,4	1,4	0,9
Chronic liver diseases	7,3		0,0					2,2				0,3		0,2		85,9									
Diseases of the digestive system (other)			0,4	0,5			1,3		0,3		0,5	0,3		0,4	0,4	1,3	81,0	1,0				3,5	1,1	0,1	0,3
Diseases of the genitourinary system		0,3	0,1			0,7					1,4			0,3	0,4		0,5	83,7				4,2		0,7	0,9
Congenital malformations, deformations and chromosomal abnormalities														0,1			0,1	0,3	92,4						
Symptoms, signs and abnormal clinical and laboratory finding								0,3		0,9	0,1			0,1			0,1	0,3		89,8	0,1	2,1			
Trasport accidents										5,3				0,0	0,1						96,7		1,3	0,8	
Medical procedures and therapies		6,2	0,8	2,1	4,9	2,2	1,0	0,5	1,5	1,8	1,0	1,3	2,5	1,4	1,2	0,3	5,1	6,7	1,5	4,6	0,1	58,3	0,7	2,6	2,4
External causes (other)	3,1	0,3	0,2		3,7	0,7	0,6	0,3	0,3	3,5	1,0			1,0	1,3	0,7	0,8	2,0		0,9	2,9	2,1	93,8	10,1	7,0
sequele		2,3	0,0			0,7	0,4		1,2		1,1	0,3	0,5	1,6	0,7		1,0	1,0				0,7	0,2	70,2	0,6
Other		0,3	0,1	0,9			1,5	0,3	1,2	5,3	1,1			0,2	0,9		0,4		1,5	1,9		6,3	0,3	0,5	74,8

* The reported values are column percent. On the diagonal there is the overall agreement at group level

Figure 1 – Agreement at 4 digit by trainee, for specific underlying causes of death*



* Continuos line represents the agreement by trainee (cod1-cod6) for each cause of death, dotted line represents the overall agreement by cause

3.1 Medical procedures

A special analysis is carried out on certificates containing medical procedures (table 5). For this group of certificates the level of agreement is 74.6% at 4 digit, a lower value compared to the agreement found for the other certificates (78.8%). This confirms the major complexity of surgical certificates and the difficulty in applying the coding rules. In fact, for these cases the trainees are subjected to a greater interpretation and subjectivity.

While the total certificates show an increase of the agreement over time, the medical procedures certificates have an agreement that raises gradually until the 13th week of mentoring period, reaching 77.8%, but then it decreases until 72.3%, just one point percent more than the first weeks. Moreover the presence of medical procedures on the death certificate increases the inter-coder variability: the range of variation of the agreement (67%-78%) is wider than the range observed for the entire set of coded certificates (76%-80%).

Table 5 – Agreement in certificates with mention of medical procedures and comparison with other certificates

	Number of certificates	Agreement %	IC95%	
			inf	sup
Certificates mentioning medical procedures	4,686	74.6	73.3	75.8
Other certificates	19,614	78.8	78.2	79.3

Table 6 – Agreement in certificates with mention of medical procedures by mentoring week and by trainee

	Number of certificates	Agreement %	IC95%	
			inf	sup
Week of mentoring		Overall		
1-4	508	71.3	67.3	75.2
5-8	868	75.7	72.8	78.5
9-13	1,488	77.8	75.7	79.9
13-17	1,822	72.3	70.3	74.4
Trainee		By trainee		
Trainee 1	781	78.1	75.2	81.0
Trainee 2	781	73.2	70.1	76.3
Trainee 3	781	76.3	73.3	79.3
Trainee 4	781	66.8	63.5	70.1
Trainee 5	781	76.3	73.3	79.3
Trainee 6	781	76.7	73.7	79.7

4. Discussion

In this study, an agreement of UCD selection equal to 78.0% at ICD-10 4 digit level and 82.3% at 3 digit was found. Despite the difficulties to compare the studies of reliability because of different applied methodologies, our results fit in with other works of coding reliability and in comparison with the other countries, we perform on average. Nevertheless, comparisons with other studies are impaired by different coding practices among countries. Some studies refer to settings where the coding is performed manually for all deaths (Harteloh et. al 2010). In this situation an agreement of 88-90% was found. Although this figure appears higher than what observed for Italy, it is necessary to take into account that the present study is based on the cases rejected by automated coding, i.e. on the most complex cases. Studies conducted in settings comparable with the Italian show an agreement of 80.9% at 3 digit level (Lu T.H. et al., 2000). On the other hand the objective of the study was to evaluate the importance of a deep training course for better mortality statistics and not to evaluate the reliability of cause-of-death data.

Actually it is not possible to reach a complete agreement between coders due to many factors such as inappropriate completion of certificate by the certifying physician, personal interpretation of medical terms, complexity of ICD and different interpretation of coding rules.

In this work the degree of learning ICD-10 rules was evaluated in order to assure the agreement of the trainees with the UCD attributed by senior coders. Prior to introduce the trainees in the routine of national cause of death coding, we wanted to verify the comparability of coding with the senior coders in order to avoid discontinuity in data series. The need of reaching comparability with previous figures is the reason for choosing the coding performed by senior coders as gold standard.

During the mentoring period, we observed an increment of certificates coded per day (from 18 to 96), an increase of the overall agreement (from 70.8% to 78.4% at 4 digit) and a decrease of variability among trainees (from 6% to 4%). The improvement of trainee performance was due to a major coder experience, but even to regular didactic interventions in order to clarify coding doubts.

During the first 8 weeks the overall agreement between trainees and senior coders increased until a maximum of 80%. Then, from the 9th to the 17th week there was a slight decrease of the agreement (78.4%). This can be explained because some of the most problematic certificates were left stand-by and discussed at the end of the mentoring period. These certificates usually correspond to those not properly completed by the certifying physicians. Especially when the conditions are misspelled or not properly reported, the coder subjectivity plays an important role in coding. Hence for these certificates, a greater value of the variability among coders is expected. The variability of these cases is not related to the coder's training but to the poor completion of some certificates.

Similarly to other studies, the source of major discussion during the didactic interventions was on the choice of different code for equivocal terms, inappropriate judgment of casual relationship and incorrect interpretation of selection rule 3 and modification rules.

According to our results, some UCD categories need a particular attention and a continuous monitoring. The causes most subjected to variability are rheumatic heart diseases, sequelae, infectious diseases (especially viral hepatitis) and chronic liver conditions.

An additional study was carried out on the certificates mentioning medical procedures and therapies. All these certificates are coded manually, hence they significantly contribute to the quota of certificates manually coded: these certificates account for 3.6% of all deaths and represent approximately the 19% of the total rejected certificates. Coding these certificates is

not easy especially because they are often not correctly completed (e.g. the reason for surgery is omitted) and consequently coders have difficulties in attributing the correct UCD.

The overall agreement for the certificates mentioning medical procedures was lower compared to those not mentioning them (74.6% vs 78.8%). This clearly reflects the higher difficulty in UCD selection for these cases. This was also confirmed by the inter-coder variability that was greater than the one observed for the entire set of coded certificates. Moreover we observed a decrease in the agreement over time (from 77% to 72%). This was due to the pileup of the most complicated cases in the last weeks, as discussed above, but also to some erroneous coding practices that had affected the gold standard, identified during the didactic interventions. An important feed-back we had from this experience was the revision of some coding practices for medical procedures resulting in an improved specificity and quality of UCD coding.

Finally, the small number of subject participating to the study might be considered a limitation to the study, nevertheless all the measures are provided with confidence intervals and show robust results.

Other confounding factors, such as demographic and/or social characteristics of the trainees may have had an impact to the results. However we did not find any differences by gender (3 out of 6 where males) and all of the students had an university attainment.

Moreover our training course has been provided only to those that were afterwards enrolled in the official cause-of-death coding for Italy. By our point of view this is the strength of the study because it reflects a real case and not a theoretical investigation.

5 Conclusion

The reliability of cause of death coding is a hot topic in health statistics. The coding process has to be independent of the coding person, time and space. Nevertheless in practice the coding has an intrinsic variability and may influence trends of mortality statistics, not always allowing proper statistical comparison among countries or different periods. In Italy, 80% of death certificates are coded by the automated system that avoids the variability of coding; the remain 20% is coded manually by senior coders.

This work highlights the complexity of the coding process and how the coding variability can be reduced by appropriate training courses. To avoid bias in mortality official statistics, it is necessary that an in-depth know how on mortality cause coding is achieved by a single coder before he/she contributes to the statistical data processing.

After the training period, the percentage of coding agreement is 78% at 4 digit level. This agreement achieved is the same observed for other countries that release mortality data. Assuming that automatically coded certificates are not affected by variability (100% agreement), the final agreement estimated for Italy is 96%, calculated as the weighted average of automated and manual coding agreement.

Moreover, the study points out that a percentage of coding variability persists even after the coding course. This is due to an incorrect completion of death certificates by physicians and to ambiguities and weaknesses of classifications and coding rules that leaves room to the coders' personal interpretation.

We can assert that the CoD statistics are reliable in Italy with regard to the coding process as it is centrally managed and revised. Although further improvement on the reliability of coding can be achieved by clearer instructions on the ICD-10 coding rules and by improving the completion of death certificates by the physicians.

Acknowledgment

The training course was organized with the assistance of Advanced School for statistics and socio-economic analyses (SAES) especially with the collaboration of Tiziana Carrino and Antonio Ottaiano. The Authors would like to thank very much the teachers of the training course: Gennaro di Fraia, Stefano Marchetti, Marilena Pappagallo, Paola Rocchi; and the trainees which contributed to the coding: Gianfranco Alicandro, Annarita Mayer, Alessandro Mistretta, Simone Navarra, Chiara Orsi.