Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220555

Preliminary Validation of a Rule-Based System for Mortality Coding Using ICD-11

Mihai Horia POPESCU[a], Can CELIK[b], Vincenzo DELLA MEA[a] and Robert JAKOB[b]
[a] *University of Udine, Udine, Italy*
[b] *World Health Organization, Geneva, Switzerland*

Abstract. A crucial process for world-level mortality statistics is the capability to identify the underlying cause of death from death certificates. Currently such certificates are coded using ICD-10. The selection of the underlying cause is done by means of semi-automated rule-based systems. However, starting from 2022, countries should begin to adopt ICD-11, for which no system is already available. The present paper describes the architecture of a novel system for automated UC selection, with classification-independent rules, and its preliminary validation on two sets of death certificates coded with ICD-10 and ICD-11.

Keywords. ICD-11, Cause of Death, medical classifications

1. Introduction

Comparison of mortality statistics is generally done by age and sex, but the so-called underlying cause of death (UC) is the most important information used for such comparison. This is defined by the WHO (2010) as *"I (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."* [1]. For each death, the UC is selected from the causal chain of events reported by a physician on the death certificate through the application of specific rules. The death certificates are collected and coded according to a standard methodology defined by the World Health Organisation (WHO) in line with the International Statistical Classification of Diseases and Related Health Problems (ICD) [1,2].

In many nations throughout the world, the coding of the death certificate conditions and the selection of the UC is still done manually. Automated coding systems, on the other hand, have been available since the 1970s, and an increasing number of nations are willing to transition from manual to automated coding.

The available automated systems that are supporting the UC selection are mainly Iris [3,4] and ACME (Automatic Classification of Medical Entry) [5,6] which support the ICD-10 classification, while for ICD-11 there are still no available systems.

ICD has been for more than a century the main basis for comparable statistics on causes of death and non-fatal disease. The 10th revision (ICD-10) was released nearly 30 years ago, which served a variety of functions in at least 120 countries and it has been translated into 43 languages. The 11th revision was adopted by the 72nd World Health Assembly in May 2019 [2]. ICD-11 is not just an extension of categories compared to ICD-10: is a different and more powerful health information system, implemented in modern information technology infrastructures, and flexible enough for future

modification and use with other classifications and terminologies [7]. Instead of the books that represented the official release until ICD-10, ICD-11 is released in form of technological tools like ICD-11 API (Application Program Interface) [8], the Coding Tool [9], and the ICD Field Implementation Tool [10].

In the crucial transition from ICD-10 to ICD-11, any UC selection system will initially suffer of a lack of available datasets already coded with ICD-11. On the other side, the abstract selection algorithm is almost the same for both classifications.

The present paper describes the architecture of a novel system for automated UC selection, with classification-independent rules, and its preliminary validation on two sets of death certificates coded with ICD-10 and ICD-11.

2. Methods

Since the abstract selection algorithm is almost the same for ICD-10 and ICD-11, with differences related to the concrete codes involved, the main requirement for the proposed system is to be classification independent, which means, have a way to separate the selection rules from the actual codes involved. A secondary yet important requirement is the possibility to integrate with the current ICD-11 platform and tools, which in turn are designed for easy integration with third party software. Finally, rules should be easily editable by domain experts.

For the rule-based system we identified two separate modules, one for the implementation of the rule engine, and one for the implementation of the code sets, which in turn could be based on ICD-10, ICD-11 or even an ICD-10 subset called the Start-Up Mortality List (ICD-10-SMoL). The rule engine is implementing the algorithm described in the reference guide [2] described in the sections 2.19-2.20. Selecting the underlying cause of death involves two separate steps. First is it needed to identify the starting point of the sequence of conditions, then to modify the starting point, if any of the modification instructions apply. An example of rule is:

Do not accept Angina pectoris (BA40) and Chronic ischaemic heart disease (BA50-BA5Z) as due to a neoplasm.

The rules format is quite complex, but a simplified version can be viewed as:

```
NAME: Rejected Sequences - Certain ischaemic heart disease due to other
   condition
CONDITION 1: "Certain ischaemic heart disease"
BINARY OPERATION MATCH: "DUE TO"
CONDITION 2: "Neoplasm"
SELECT: "CONDITION 1"
```

This rule is used to select the new tentative UC when Angina pectoris or a Chronic ischaemic heart disease condition it is found to be due to a Neoplasm, and the selected UC is from the condition 1, which is the Certain ischaemic heart disease.

Differently from the other mortality coding systems, that treat codes as “ranges”, described only in terms of leaves of the hierarchical tree, we want to exploit the hierarchy to express the code sets at the highest abstraction level possible.

An example of code set is:

Chronic ischaemic heart disease: which has the range of *BA50-BA5Z* but can be specified in the code set as id *“http://id.who.int/icd/entity/1221742343”* that include all the Chronic ischaemic heart disease conditions.

These are the definitions (not in their JSON syntax for the sake of brevity), of the code sets for ICD-10 and ICD-11 used in the rule above, where they are mentioned by name and not by code:

```
ICD-10: "Angina pectoris" Include "I20", "Chronic ischaemic heart
   disease" Include "I25", "Neoplasm" Include "II", "Certain ischaemic
   heart disease" Include Categories ("Angina pectoris", "Chronic
   ischaemic heart disease").
ICD-11: "Angina pectoris" Include
   "http://id.who.int/icd/entity/718946808", "Chronic ischaemic heart
   disease" Include "http://id.who.int/icd/entity/1221742343", "Neoplasm"
   Include "http://id.who.int/icd/entity/1630407678", "Certain ischaemic
   heart disease" Include Categories ("Angina pectoris", "Chronic
   ischaemic heart disease").
```

2.1. Validation

As already mentioned, validation is an issue because of the lack of certificates dataset coded in ICD-11. However, our classification independent approach allows at least to validate the abstract rules on ICD-10 coded certificates. Thus, we based our validation on a dataset of death certificates coded in ICD-10 from the Center for Disease Control and Prevention (CDC) for the year 2018, plus a small dataset of certificates manually coded in ICD-11, developed ad-hoc for this work. In both datasets, the UC predicted by the system has been compared with the ground truth.

Accuracy has been then calculated, which can be computed as:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (1)$$

3. Results

3.1. The system

A library has been developed to implement the rule-based system, where the related code set has been partially implemented for two ICD classifications (ICD-11 and ICD-10) and ICD-10-SMoL. The library was developed in *.dotnet* framework, which can be deployed in most of the environments. To make use of the library, two applications have been developed: a console-based and a web-based application. The web application back-end was implemented with support of extended technologies like Application Program Interface (API) which gives the opportunity to support implementation of different devices applications.

The rule-based system is implemented in two separate modules, one which specifies the rules in a JSON format, and an algorithm implemented in *C#* that interprets the rules and execute them. Since the rules are defined separately, give the freedom to modify the algorithm logic without changing the associated programming code. This is also the reason why the actual rules do not have any knowledge of the classification behind, but just implement the logic of the algorithm. In order to extend the system with a new classification, a few classes would be need to be extended in the library to support some classification functionalities and implement the related code sets.

The code sets are implemented in multiple JSON files, which the rules refer to for its evaluation, reason why the same code set should be implemented for each classification.

Currently the system fully implements the algorithm, while the code set is partially implemented with similar percentage of completeness for ICD-10 and ICD-11. 18 out of 38 selection rules are fully implemented, where the others need further domain expert support. Near to 95% of the modification rules are implemented, where the remaining need again expert intervention.

Table 1. Effectiveness scores for the analysis of the proposed system, for the analysis made in the Netherlands and for ML approach.

No. certificates	Certificates dataset	System used to select UC	Rejected certificates	Accuracy
2.846.305	ICD-10 dataset	Proposed solution	8.2%	78%
1.248	ICD-11 dataset	Proposed solution	11.6%	62.8%
134.262	ICD-11 Netherlands	Iris [11]	31.5%	78%
400.000	ICD-10 CDC	ML [12]	0%	98.75%

In Table 1 we can observe the effectiveness scores that the proposed system obtained for the ICD-10 dataset and for the ICD-11 dataset. Certificates may be rejected for wrong codes, classification version mismatch, ad in the case of ICD-11, wrong postcoordination. Removing the rejected certificates from the analysis, the system was able to correctly select the UC with an accuracy of 78%. On the ICD-11 dataset, accuracy is 62.8%. In the latter case, the rejected certificates mostly depend on a version mismatch, due to the rapid evolution of ICD-11.

4. Discussion

The preliminary validation of the proposed mortality coding system has been carried out on a large dataset of ICD-10 coded death certificates and on a very small dataset of ICD-11 coded certificates. Although not directly comparable due to the different dataset, for ICD-10 coded certificates preliminary results show an accuracy very close to the one found in a similar study on IRIS [11], while for the ICD-11 data set accuracy is still lower. However, the ICD-11 dataset, which only contains 1248 certificates, might not fully represent the real-world distribution of causes of death, over-representing less frequent cases. On the other side, such accuracy has been obtained with an incomplete set of rules.

A consideration should be done also for systems based on Machine Learning techniques, which have been show to outperform rule-based systems on ICD-10 coded certificates reaching an accuracy of 98.75% [12,13]. Although the results reached, those systems cannot be implemented yet for the new 11th revision. ML has great capabilities and can give great support for this purpose, but to perform they need great quantity of data for the training and the dataset need to be of highest quality, which is not always possible to ensure in an early implementation.

5. Conclusions

In this paper we have presented the first rule-based system with the capability to select the underlying cause of death for the ICD 11th revision, while still being able to work

with the previous revision of the classification. The results seem very promising, and this gives the possibility to mortality coding in the early stage of ICD-11 adoption for statistics, while looking forward for the full implementation of the rules and code sets, and enable a complete validation. A crucial future work, needed for validating this system as well as other future systems, is the development of a data set of ICD-11 coded death certificates, possibly developed by coders in different countries, to cover the inter-country variability in cause of death distribution and coding style.

6. Acknowledgements

We thank the Mexican WHO-FIC Collaborating Center for having developed the ICD-11 coded data set. We thank Robert N. Anderson at the Mortality Statistics Branch, U.S. National Center for Health Statistics, for having provided the ICD-10 coded data set.

References

[1] World Health Organization - WHO. International Statistical Classification of Diseases and Related Health Problems. 10th Revision, vol 2, Fifth edition 2016. Geneva, Switzerland: WHO. Retrieved from https://icd.who.int/browse10/Content/statichtml/ICD10Volume2_en_2016.pdf
[2] World Health Organization. ICD-11 revision. https://icd.who.int/en. Accessed 13 Gen 2022
[3] Iris Institute. Official website. www.iris-institute.org
[4] Pavillon G, Johansson LA, Glenn D, Weber S, Witting B, and Notzon S. 2007. Iris: A Language Independent Coding System For Mortality Data. In WHO-FIC. Annual Meeting. Trieste, Italy, 2007.
[5] Centre for Disease Control and Prevention - CDC, National Centre for Health Statistics – NCHS. National Vital Statistics System. 2021. Instruction manuals. ICD-10 ACME Decision Tables for Classifying Underlying Causes of Death.
[6] Lu TH. Using ACME (Automatic Classification of Medical Entry) software to monitor and improve the quality of cause of death statistics. J Epidemiol Community Health. 2003 Jun; 57(6): 470–471.
[7] Harrison JE, Weber S, Jakob R, Chute CG. ICD-11: an international classification of diseases for the twenty-first century. BMC Med Inform Decis Mak 21, 206 (2021).
[8] Mugisha M, Byiringiro JB, Uwase M, Abizeyimana T, Ndikubwimana B, Karema N, Kayiganwa M, Tran Ngoc C, Kostanjsek N, Celik C, Marc D, Twizere C, Tumusiime D. Integration of International Classification of Diseases Version 11 Application Program Interface (API) in the Rwandan Electronic Medical Records (openMRS): Findings from Two District Hospitals in Rwanda. Stud Health Technol Inform. 2020 Jun 26;272:280-283.
[9] ICD-11 Coding Tool. Mortality and Morbidity Statistics (MMS). https://icd.who.int/ct11/icd11_mms/en/release Accessed 13 Gen 2022
[10] Donada M, Kostanjsek N, Della Mea V, Celik C, Jakob R. Piloting a Collaborative Web-Based System for Testing ICD-11. Stud Health Technol Inform. 2017;235:466-470.
[11] Harteloh P. The implementation of an automated coding system for cause-of-death statistics. Informatics for Health and Social Care. vol 45, no 1, pp 1-14. 2020.
[12] Della Mea V, Popescu MH, Roitero K. Underlying Cause of Death Identification from Death Certificates using Reverse Coding to Text and a NLP Based Deep Learning Approach. Informatics in Medicine Unlocked, 21, 100456 (2020)
[13] Falissard L, Morgand C, Roussel S, Imbaud C, Ghosn W, Bounebache K, Rey G. A Deep Artificial Neural Network-Based Model for Prediction of Underlying Cause of Death From Death Certificates: Algorithm Development and Validation. JMIR Med Inform 2020;8(4):e17125