Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200114

A Semi-Automated Approach for Multilingual Terminology Matching: Mapping the French Version of the ICD-10 to the ICD-10 CM

Emmanuelle SYLVESTRE[abc1], Guillaume BOUZILLÉ[ab], Michael McDUFFIE[d], Emmanuel CHAZARD[e] Paul AVILLACH[d], Marc CUGGIA[ab]
[a] *INSERM, LTSI UMR 1099, F-35000, Rennes France*
[b] *CHU Rennes, Centre de Données Cliniques, F-35000, Rennes France*
[c] *CHU Martinique, Centre de Données Cliniques, F-97200, Martinique France*
[d] *Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115 USA.*
[e] *Université de Lille, CHU Lille, CERIM EA2694, F-59000 Lille, France.*

Abstract. The aim of this study was to develop a simple method to map the French International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10) with the International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10 CM). We sought to map these terminologies forward (ICD-10 to ICD-10 CM) and backward (ICD-10 CM to ICD-10) and to assess the accuracy of these two mappings. We used several terminology resources such as the Unified Medical Language System (UMLS) Metathesaurus, Bioportal, the latest version available of the French ICD-10 and several official mapping files between different versions of the ICD-10. We first retrieved existing partial mapping between the ICD-10 and the ICD-10 CM. Then, we automatically matched the ICD-10 with the ICD-10-CM, using our different reference mapping files. Finally, we used manual review and natural language processing (NLP) to match labels between the two terminologies. We assessed the accuracy of both methods with a manual review of a random dataset from the results files. The overall matching was between 94.2 and 100%. The backward mapping was better than the forward one, especially regarding exact matches. In both cases, the NLP step was highly accurate. When there are no available experts from the ontology or NLP fields for multi-lingual ontology matching, this simple approach enables secondary reuse of Electronic Health Records (EHR) and billing data for research purposes in an international context.

Keywords. ICD-10, Clinical terminologies, Interoperability, Multilingual matching

1. Introduction

The International Statistical Classification of Diseases and Related Health Problems, 10th revision (WHO-ICD-10) one of the most popular terminologies used in around the world.

[1] Corresponding Author, *Faculté de médecine, Université Rennes 1, 2 Avenue du Professeur Léon Bernard 35043 Rennes Cedex 9, France*; E-mail: emmasyl@gmail.com.

It is a standard diagnostic terminology created and maintained by the World Health Organization (WHO) since 1990 for diagnostic coding[1]. The French healthcare system uses a French version of the WHO-ICD-10 (Classification Internationale des Maladies, 10e version, CIM-10) since 1997[2], while the United States created and implanted their own adaptation of the terminology (International Classification of Diseases, 10th Revision, Clinical Modification, ICD-10-CM) on October, 1st, 2015[3]. There are far more codes in the ICD-10 CM than in the CIM-10, even though they share the same common denominator: the WHO-ICD-10[3].

Multilingual ontology matching is the process of finding correspondences between ontologies of different languages to allow them to interoperate[4]. This enables secondary reuse of Electronic Health Records (EHR) and billing data from different healthcare systems for research purposes, especially if data from the United States is involved in the study. We can use two main strategies for multilingual ontology matching: direct and indirect alignment. The direct alignment is translation-based and uses external resources to help with translation, while the indirect alignment uses intermediary mappings between the source and target ontologies. Furthermore, mapping two ontologies can be an automated or manual process. Manual mapping is still the prevalent choice for ontology matching, but necessitates a large team of experts, is time-consuming, and is prone to errors[5]. On the other hand, automated approaches use public terminology resources such as the Unified Medical Language System (UMLS)[4] or Bioportal[6] but those sources are extremely incomplete outside of the English speaking world[7]. Therefore, when the purpose of a study is not the mapping itself but a necessary step to join databases, it should be possible to overcome the semantic interoperability issue by combining different automated matching techniques to conduct the study, even with limited resources or experts from the ontology matching field.

The aim of this study was to develop and evaluate a simple method to link the French ICD-10 (or any version of the WHO ICD-10) with the ICD-10 CM. We sought to map these terminologies forward (ICD-10 to ICD-10 CM) and backward (ICD-10 CM to ICD-10).

2. Methods

2.1. Terminology resources

Since there were no direct mapping files for our study, we used all the intermediate mapping files available online. We used four data sources: i) the UMLS Metathesaurus, which integrates and assigns a unique identifier to synonymous concepts from several standard biomedical technologies (including ICD-10 CM, WHO-ICD-10 and a 1998 version of the CIM-10)[4], ii) Bioportal, which is a comprehensive repository of standard terminologies created by the National Center for Biomedical Ontology (NCBO), with an ontology alignment tool, called Lexical OWL Ontology Matcher (LOOM)[6], iii) the latest version of the CIM-10, which is mainly a translation of the WHO-ICD-10 with more children, and is publicly available on the national billing agency website (Agence Technique de l'Information sur l'Hospitalisation, ATIH)[2] and iv) different existing mapping files with forward and backward mapping such as: the General Equivalent Mapping (GEM) Files from the National Center for Health Statistics (NCHS)[8], the New-Zealand mapping files from the Ministry of Health (ICD-10 Australian Modification and ICD-9 Australian Modification)[9] and the mapping files between the

ICD-10 AM and ICD-10, from the Australian Consortium for Classification Development (ACCD)[10].

2.2. Mapping method

We decided to map only the first 4-character codes of the CIM-10, because the WHO-ICD-10 is mainly 4-character codes, except for chapters XIII, XIX, XX, which means that after this position, there was a bigger risk of mapping codes with a different signification between the two languages.
Our strategy used three main steps (Figure 1). First, we used the NCBO API to retrieve the mapping between the ICD-10 CM and the WHO-ICD-10 and the partial mapping between the CIM-10 and the WHO-ICD-10 from the UMLS Metathesaurus and LOOM algorithm, which is available through the NCBO API. Then, we matched each terminology using the different mapping files mentioned above. And finally, we used manual review and natural language processing (NLP) to recognize labels between the two terminologies with custom R scripts. The NLP process was only used on unmatched codes after the two first steps. For each set, one native-speaking investigator from each language extracted the two or three main words of the ICD or CIM label, then they built together a translation dictionary and used rules-based NLP to match the labels and codes. For the remaining unmatched codes, we mapped them to their three-character codes parent. After all these steps, we considered the code unmatched if we could not find an exact or approximate match.

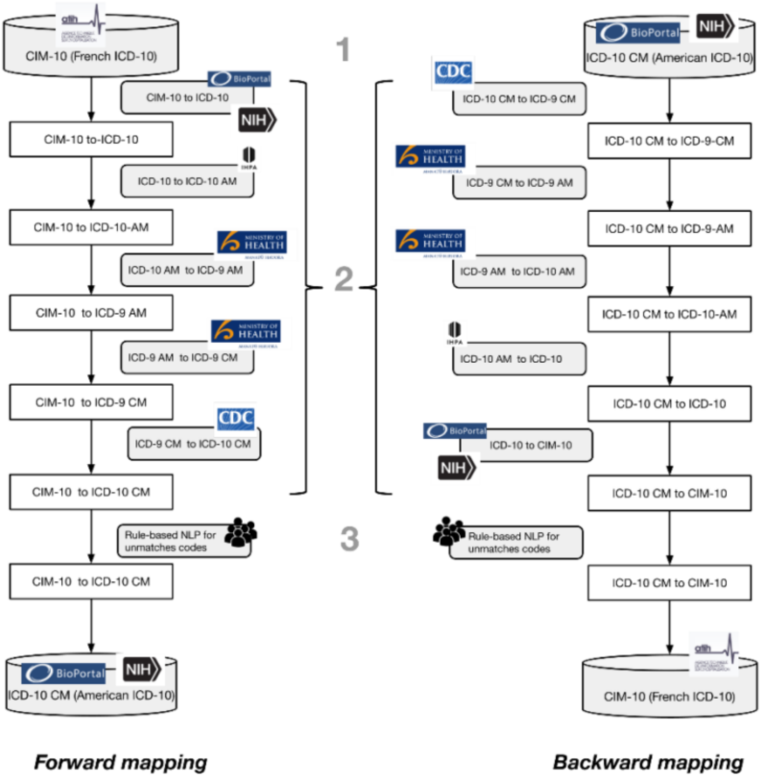

Figure 1. Forward and backward mapping methods

Since the automated mapping was based on official mapping files, we did not review those matched codes. However, we reviewed manually all NLP-based matches before confirming their match or un-match status. Uncertain pairs were reviewed again by two other medical investigators.

3. Results

The ICD-10 CM included 91,737 codes and the French ICD-10 included 39,928 codes, including the 3-charcters codes parents. The ICD-10 CM had 9835 (11%) of 4-characters codes while the ICD-10 had 12,345 (31%) of 4-character codes.

Among the 39,928 codes of the CIM-10, 8,477 (21.2%) were exact matches to the ICD-10 CM and 29,131 (73 %) were partial matches. 264 of those partial matches were based on the NLP-based step. There were 2,320 (5.8%) missing codes.

Among the 91,737 codes of the ICD-10 CM, 9,082 (9.9%) were exact matches and 82,655 (90.1 %) partial matches with no unmatched codes. 226 of those partial matches were based on the NLP-based step. There were no missing codes.

All the NLP-based matches were true positives with no false negatives. Overall, the backward mapping was 94.2%, while the forward mapping was 100%. (Table 1)

Table 1. Characteristics of all matches after the forward and backward mappings

Mapping	Exact Match	Partial Match*				No match
		>4-characters code to one 4-character code	>4-characters code to more than one 4-characters code	4-characters code to more than one 4-characters code	4-characters codes (or more) to the 3-characters parent	
ICD-10 to ICD-10 CM	8,477 (21.2%)	4,832 (12.1%)	23,345 (58.5%)	932 (2.3%)	22 (0.1%)	2,320 (5.8%)
ICD-10CM to ICD-10	9,082 (9.9%)	41,518 (45.3%)	33,893 (36.9%)	1,797 (2%)	5,447 (5.9%)	0 (0%)

*The partial match includes the codes matched using the Natural Language Processing (NLP) step

4. Discussion

Our method showed a very high matching score, especially regarding the backward mapping (ICD-10 CM to ICD-10) with 100% of codes matched. However, the exact match was far better in the forward mapping (21% versus 10% in the backward). The manual evaluation confirmed the accuracy of the rules-based NLP algorithm.

Our study has several limitations. First, we only tested the method with two languages (French and English). However, we only had native speaking experts in French and English available for the manual review and most countries use either the ICD-10 CM, or a fairly close version of the WHO ICD-10 for reimbursement and billing[1]. Therefore, we knew that if our mapping process was accurate it would be relatively easy to adapt it to other languages. Second, our NLP was rather basic because we wanted first and foremost to use already existing reference files. Previous studies[11,12] showed that NLP based on modern machine-learning methods is the most pertinent method to match a diagnostic to an ICD-10 code or to translate a terminology to another language, but it is a very specific field with few experts, especially outside of the English language. The idea here was to propose an alternative when NLP specialists are not available to implement automatic translation and/or matching algorithms. Finally, most of our matches are partial. A majority of exact matches would have been ideal, especially since the ICD-10 is not always very precise regarding some diagnoses[13], but since the WHO-ICD-10 and the ICD-10-CM have a vastly different number of codes[3], this outcome was predictable.

Manual mapping for multilingual ontology alignment is still the gold standard today, but it requires several experts of the healthcare and ontology fields and is a very time-consuming work that can take years[14]. This method cannot make the same claims of precision as official manual mapping files, but it could become a fairly quick and reliable process for international studies based on secondary reuse of EHR and billing data (including legacy data coded in ICD-9 CM, thanks to the GEM files from the NCHS[8]) without any ontology or translation experts.

5. Conclusion

Our study demonstrated that semi-automated mapping based on reference mapping files (in standard format) and basic NLP could be considered for secondary reuse of EHR and billing data from different countries when there are no existing reference files. Next, we would like to replicate this study with ICD-10 in other languages and use more automated NLP resources, like the Google Translate API, to confirm the accuracy of this method. This work could also serve as a basis for semi-automated mapping of the ICD-10 to the ICD-11, once the official ICD-10 to ICD-11 mapping files are available.

References

[1] WHO | International Classification of Diseases (ICD) Information Sheet, *WHO*. (n.d.). https://www.who.int/classifications/icd/factsheet/en/.

[2] ATIH : Agence technique de l'information sur l'hospitalisation, (n.d.). http://www.atih.sante.fr/.

[3] L. Manchikanti, A.D. Kaye, V. Singh, and M.V. Boswell, The Tragedy of the Implementation of ICD-10-CM as ICD-10: Is the Cart Before the Horse or Is There a Tragic Paradox of Misinformation and Ignorance?, *Pain Physician*. **18** (2015) E485-495.

[4] US National Library of Medicine. Unified Medical Language System (UMLS): Metathesaurus, (n.d.). https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html.

[5] C.T. Dos Santos, P. Quaresma, and R. Vieira, An API for multilingual ontology matching, in: 2010.

[6] P.L. Whetzel, N.F. Noy, N.H. Shah, P.R. Alexander, C. Nyulas, T. Tudorache, and M.A. Musen, BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications, *Nucleic Acids Res.* **39** (2011) W541-545.

[7] A. Doan, J. Madhavan, P. Domingos, and A. Halevy, Ontology matching: A machine learning approach, in: Handb. Ontol., Springer, 2004: pp. 385–403.

[8] ICD - ICD-10-CM - International Classification of Diseases, Tenth Revision, Clinical Modification, (2019). https://www.cdc.gov/nchs/icd/icd10cm.htm.

[9] Mapping between ICD-10 and ICD-9 | Ministry of Health NZ, (n.d.). http://www.health.govt.nz/nz-health-statistics/data-references/mapping-tools/mapping-between-icd-10-and-icd-9.

[10] Australian Consortium for Classification Development, The international statistical classification of diseases and related health problems, tenth revision, Australian modification (ICD-10-AM/ACHI/ACS): Mapping files, (n.d.). https://ace.ihpa.gov.au/Downloads.aspx.

[11] N.D. Hailu, K.B. Cohen, and L.E. Hunter, Ontology translation: A case study on translating the Gene Ontology from English to German, *Nat. Lang. Process. Inf. Syst. Int. Conf. Appl. Nat. Lang. Inf. Syst. NLDB Revis. Pap. Int. Conf. Appl. Nat. Lang. Info*. **8455** (2014) 33–38.

[12] A. Atutxa, A.D. de Ilarraza, K. Gojenola, M. Oronoz, and O. Perez-de-Viñaspre, Interpretable deep learning to map diagnostic texts to ICD-10 codes, *Int. J. Med. Inf.* **129** (2019) 49–59.

[13] Y.M. Mesfin, A.C. Cheng, A.H.L. Tran, and J. Buttery, Positive predictive value of ICD-10 codes to detect anaphylaxis due to vaccination: A validation study, *Pharmacoepidemiol. Drug Saf.* (2019).

[14] ICD-11 Content - Content Development Roadmap - SNOMED Confluence, (n.d.). https://confluence.ihtsdotools.org/display/CDR/ICD-11+Content.