pHealth 2022
B. Blobel et al. (Eds.)

doi:10.3233/SHTI220987

Analysis of the Representation of Frequent Clinical Attributes in the Unified Medical Language System

Baris GÜNGÖR[a], Noemi DEPPENWIESE[b], Jonathan M. MANG[b] and Dennis TODDENROTH[a,1]
[a] *Medical Informatics, University Erlangen-Nuremberg, Germany*
[b] *Medical Center for Information and Communication Technology, University Hospital Erlangen, Germany*

Abstract. Mapping clinical attributes from hospital information systems to standardized terminologies may allow their scientific reuse for multicenter studies. The Unified Medical Language System (UMLS) defines synonyms in different terminologies, which could be valuable for achieving semantic interoperability between different sites. Here we aim to explore the potential relevance of UMLS concepts and associated semantic relations for widely used clinical terminologies in a German university hospital. To semi-automatically examine a sample of the 200 most frequent codes from Erlangen University Hospital for three relevant terminologies, we implemented a script that queries their UMLS representation and associated mappings via a programming interface. We found that 94% of frequent diagnostic codes were available in UMLS, and that most of these codes could be mapped to other terminologies such as SNOMED CT. We observed that all examined laboratory codes were represented in UMLS, and that various translations to other languages were available for these concepts. The classification that is most widely used in German hospital for documenting clinical procedures was not originally represented in UMLS, but external mappings to SNOMED CT allowed identifying UMLS entries for 90.5% of frequent codes. Future research could extend this investigation to other code sets and terminologies, or study the potential utility of available mappings for specific applications.

Keywords. Unified Medical Language System, Semantic Interoperability, Application Programming Interface.

Introduction

The introduction of electronic health records intent to improve the health care of individuals, and for researchers increases the possibilities to gain more knowledge about diseases and treatments. It has been proposed that realizing this potential requires that a critical mass of health care providers adopt the collection of patient data in a digital format, as well as interoperability between the involved systems [1].

Semantic interoperability between different systems can be achieved through the use of standard terminologies. If an international study attempted to accumulate clinical

[1] Corresponding Author: Dennis Toddenroth, PhD Ass. Professor; Institute for Medical Informatics, Biometry and Epidemiology, University of Erlangen, Erlangen, Germany; Email: dennis.toddenroth@fau.de.

records from participating sites for an aggregate analysis, the uniform use of a standardized terminology would allow a consistent scientific reuse by simply pooling contributed datasets. If no homogenous semantic reference is available, on the other hand, a mapping between the attributes of the participating sites would have to be set up manually, which can be laborious.

The four consortia of the German Medical Informatics Initiative are currently in the process of establishing data integration centers at university hospitals, which aim to support medical research by collecting and harmonizing patient records from routine care information systems [2]. These organizations manage data reutilization with specific provisions that consider the privacy rights of the involved patients, and ideally reuse clinical data for multiple studies, so expanding the semantic integration of frequent attributes would be particularly valuable.

The Unified Medical Language System (UMLS), developed by the U.S. National Library of Medicine, was chosen for this work because of its more than 3.25 million concepts (as of 2016) from over 220 source vocabularies (as of 2021), making it one of the richest thesauri in biomedicine [3,4]. UMLS is regularly updated and extended. It allows users to access terminology data via an application programming interface (API), which enables automated fields of application [5,6].

The goal of this study is to explore the potential relevance of UMLS concepts and associated semantic relations for frequently used clinical attributes in a German university hospital by semi-automatically analyzing their representation in UMLS.

1. Method

To analyze the availability of relevant clinical attributes in the UMLS metathesaurus, we automatically queried the provided API with lists of the most frequently documented codes for three commonly used clinical terminologies.

We focused on "The International Statistical Classification Of Diseases And Related Health Problems, 10th revision, German Modification" (ICD-10-GM), "Logical Observation Identifiers Names and Codes" (LOINC) and "German procedure classification" (*Operationen- und Prozedurenschlüssel* - OPS), based on their relevance for routine clinical documentation and on their usage at Erlangen University Hospital [7]. ICD-10-GM is the statutory classification for coding diagnoses in inpatient medical care in Germany [8], LOINC codes are used for identifying clinical observations and measurements such as laboratory tests [9], while OPS codes denote diagnostic and therapeutic medical procedures [10].

As a subsample, we focused on the 200 most frequent codes from Erlangen University Hospital for each of the three terminologies. To automatically examine the UMLS representation and potential semantic mappings, we implemented a Python script that iterates through input codes and communicates with the UMLS API. Information from the API responses is first stored and then visualized according to the following recipe:

1. The top 200 ICD-10-GM codes are provided as input.
2. For each code, the concept ID in UMLS is identified.
3. All atoms associated with this concept ID are queried and stored in a separate list.
4. This output is used to count for each terminology how many of the 200 codes have a mapping to that specific terminology.

5. For visualization, the output list from step 4 is imported into Excel and displayed in a diagram.

The same procedure is used for LOINC codes. The mapping of OPS codes required another step since we observed that UMLS did not originally contain any OPS codes. A previous research project with TrinetX, a global health research network [11,12], had produced and published a list of OPS to SNOMED CT mappings, which we used for an intermediate mapping step.

The developed script that accesses the UMLS API is available at https://github.com/brsgngr/UMLSMappingsReader.

2. Results

The procedure described above leads to the results shown in figures 1 to 3. Figure 1 illustrates the frequency distribution of available UMLS mappings for the 200 most common ICD-10-GM codes, Figure 2 shows the corresponding frequencies for the 200 most frequent LOINC codes, and Figure 3 displays the same data for the 200 most frequent OPS codes.

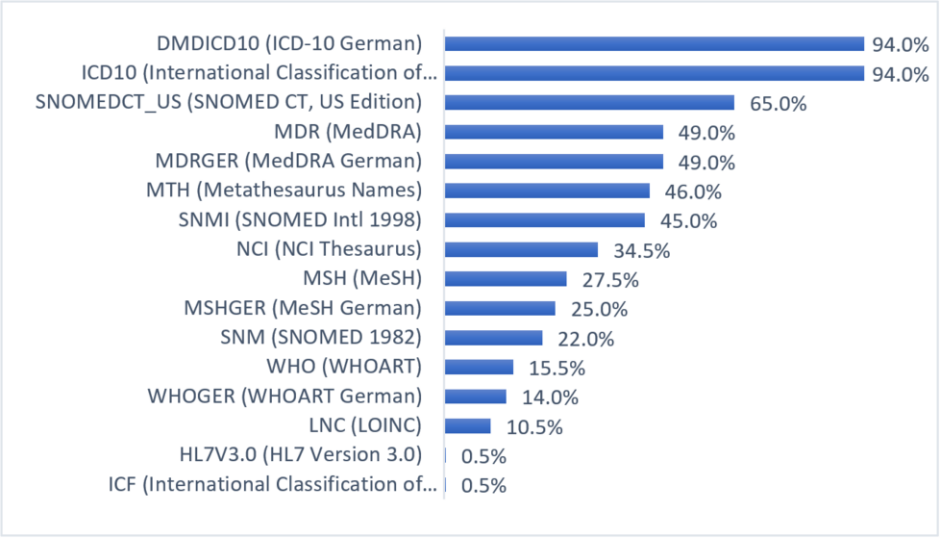

Figure 1. Frequency distribution of a subset of available UMLS mappings for the 200 most frequent ICD-10-GM codes. Out of the 200 studied ICD-10-GM codes, 94.0% were found in UMLS and 65.0% had a SNOMED CT mapping. The terminology labels are used as defined by UMLS.

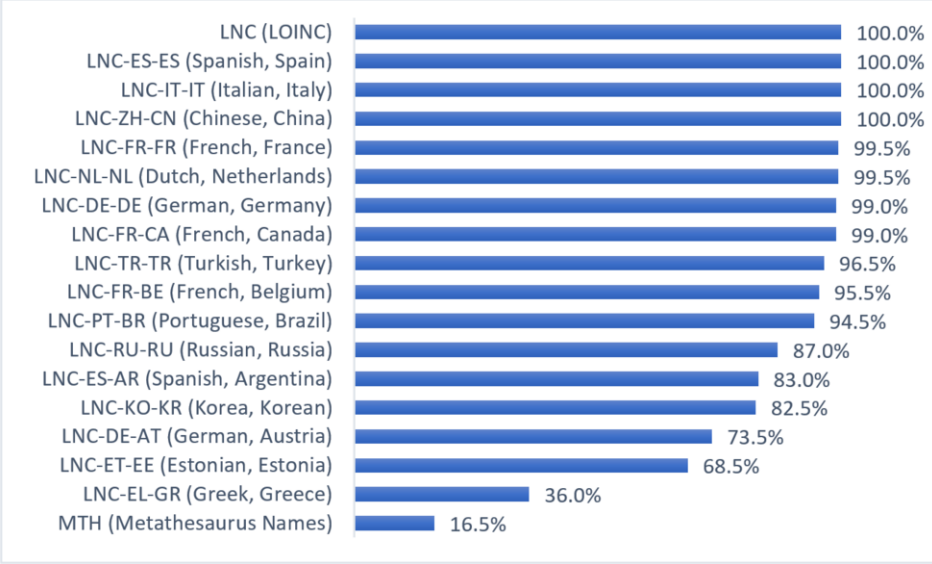

Figure 2. Frequency distribution of available UMLS mappings for the 200 most frequent LOINC codes. The terminology labels are used as defined by UMLS.

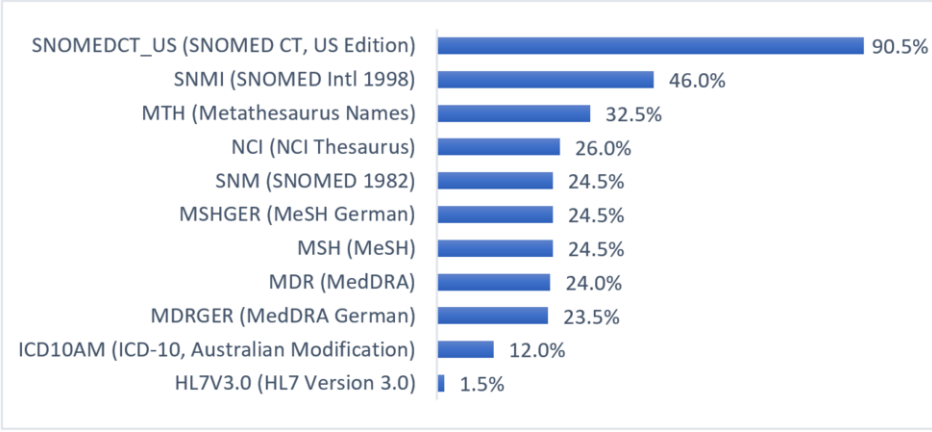

Figure 3. Frequency distribution of a subset of available UMLS mappings for the 200 most frequent OPS codes after mapping the OPS codes to SNOMED CT codes. The terminology labels are used as defined by UMLS.

The Figures can be read as follows: 65.0% of the 200 most frequent ICD-10-GM codes had a SNOMED CT mapping in UMLS.

Also, it can be stated that 94% of the most frequently used 200 ICD-10-GM codes were found in UMLS, whereas 6.0% were not found. All LOINC codes were present in UMLS. For the OPS codes, 90.5% of the codes were found in UMLS after mapping the OPS code to SNOMED CT. The remaining OPS codes could not be identified either because the OPS code could not be mapped to any SNOMED CT code or UMLS does did not contain the corresponding SNOMED CT code.

Note that Figures 1 and 3 display only a subset of mappings, because only terminologies with frequent mappings have been included in order to improve readability. Figure 2 does not include any filtering.

3. Discussion

We observed that for all three terminologies at least 91%, and for LOINC even 100% of the 200 most frequent codes could be identified in UMLS. For ICD-10-GM as well as for OPS, many mappings were found, also to terminologies that are relevant in Germany, whereas LOINC codes were mostly mapped to LOINC codes in other languages (i.e. translations).

Three aspects were observed when examining the most frequent ICD-10-GM codes. First, in UMLS, an ICD-10-GM code only exists if the corresponding ICD-10 code is present. This is the reason for the imperfect overlap of the ICD-10-GM codes with the ICD-10 codes in Figure 1. Second, it is important to note that the initial list of the 200 most frequent ICD-10-GM codes actually included 33 5-digit codes. Since ICD-10 only allows 4-digit codes, the 5-digit ICD-10-GM codes could not be found in UMLS. If the higher-level category, i.e. the 4-digit code, is used for these 33 codes, it generates the result shown in Figure 1. Third, eight codes from the second run could not be found in UMLS, because these codes were added to ICD-10 after the last update of ICD-10 in UMLS. Overall, it can be stated that treating ICD-10-GM more as an independent terminology in UMLS could enable the integration of 5-digit codes and would thus also allow a more precise mapping analysis. A feasible extension of our script would thus automatically revert to the higher-level category of a code if it is not found.

Examination of the 200 most frequent LOINC codes revealed that even though all LOINC codes were found in UMLS, a large fraction of codes was only associated with translated versions of these LOINC codes itself. This makes it difficult to draw conclusions about other terminologies. Perhaps LOINC codes could be mapped to a different terminology using other data sources and examined in UMLS using this other terminology.

Although OPS codes were not directly represented in UMLS, some semantic relations to other terminologies could be found in UMLS with the help of the TriNetX mapping. In this way, results similar in quality to ICD-10-GM were achieved. One difference, however, is that in the ICD-10-GM yielded more mappings to other relevant terminologies.

Through the mappings across terminologies and different concepts, the scientific use can be increased, further fields of knowledge can be identified, and additional conclusions can be made. An exemplary use could be to suggest publications or clinical studies appropriate to specific clinical situations based on linked codes, such as through Medical Subject Headings (MeSH Terms). A high rate of mappings between terminologies increases the potential for the different application fields.

In the future, the Python script could be run with additional codes in order to gain insight into the available UMLS mappings for a different set of codes. It could be extended to also support other terminologies, or underpin prototypical applications that leverage such dynamic semantic mapping, such as the use case of suggesting relevant publications based on mapping clinical codes to synonymous MeSH codes.

Acknowledgements

This work was conducted within the MIRACUM consortium. MIRACUM is funded by the German Ministry for Education and Research (BMBF) [funding number FKZ 01ZZ1801A].

References

[1] Kanter AS, Wang AY, Masarie FE, Naeymi-Rad F, and Safran C. Interface Terminologies: Bridging the Gap between Theory and Reality for Africa. Studies in Health Technology and Informatics 2008; 136: 27-32. DOI: 10.3233/978-1-58603-864-9-27.

[2] Prokosch HU, Acker T, Bernarding J, Binder H, Boeker M, Boerries M, Daumke P, Ganslandt T, Hesser J, Höning G, Neumaier M, Marquardt K, Renz H, Rothkötter HJ, Schade-Brittinger C, Schmücker P, Schüttler J, Sedlmayr M, Serve H, Sohrabi K, and Storf H. MIRACUM: Medical Informatics in Research and Care in University Medicine. Methods of Information in Medicine 2018 Jul;57(S 01): e82-e91. doi:10.3414/ME17-02-0025.

[3] Lu CJ, Tormey D, McCreedy L, and Browne AC. Enhanced lexsynonym acquisition for effective UMLS concept mapping. Studies in Health Technology and Informatics 2017; 245: 501-505. doi:10.3233/978-1-61499-830-3-501.

[4] Declerck G, Souvignet J, Rodrigues JM, and Jaulent MC. Automatic annotation of ICD-to-MedDRA mappings with SKOS predicates. Studies in Health Technology and Informatics 2014; 205: 1013-7. doi:10.3233/978-1-61499-432-9-1013.

[5] Patel CO and Cimino JJ. A scale-free network view of the umls to learn terminology translations. Studies in Health Technology and Informatics 2007; 129(Pt 1): 689-93.

[6] Raje S, and Bodenreider O. Interoperability of disease concepts in clinical and research ontologies: Contrasting coverage and structure in the disease ontology and SNOMED CT. Studies in Health Technology and Informatics 2017; 245: 925-929. doi:10.3233/978-1-61499-830-3-925.

[7] BfArM - Kodiersysteme, (n.d.). https://www.bfarm.de/DE/Kodiersysteme/_node.html (accessed February 24, 2022).

[8] BfArM - ICD-10-GM, (n.d.). https://www.bfarm.de/EN/Code-systems/Classifications/ICD/ICD-10-GM/_node.html (accessed February 27, 2022).

[9] BfArM - LOINC, (n.d.). https://www.bfarm.de/DE/Kodiersysteme/Terminologien/LOINC-UCUM/LOINC-und-RELMA/_node.html (accessed February 27, 2022).

[10] BfArM - OPS, (n.d.). https://www.bfarm.de/EN/Code-systems/Classifications/OPS-ICHI/OPS/_node.html (accessed February 27, 2022).

[11] Schulz S, Steffel J, Polster P, Palchuk M, Daumke P. Aligning an administrative procedure coding system with SNOMED CT, CEUR Workshop Proceedings 2019; 2518.

[12] Download - open.trinetx, (n.d.). https://open.trinetx.com/download/ (accessed February 24, 2022).