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

doi:10.3233/SHTI220464

Mapping of ICD-O Tuples to OncoTree Codes Using SNOMED CT Post-Coordination

Tessa OHLSEN[a,1], Valerie KRUSE[b], Rosemarie KRUPAR[c], Alexandra BANACH[a], Josef INGENERF[a,d] and Cora DRENKHAHN[d]
[a] *Institute for Medical Informatics, University of Lübeck, Lübeck, Germany*
[b] *Clinic for Hematology and Oncology, UKSH, Lübeck, Germany*
[c] *Pathology of the Research Center Borstel, Leibniz Lung Center, Borstel, Germany*
[d] *IT Center for Clinical Research, University of Lübeck, Lübeck, Germany*

Abstract. Around 500,000 oncological diseases are diagnosed in Germany every year which are documented using the International Classification of Diseases for Oncology (ICD-O). Apart from this, another classification for oncology, OncoTree, is often used for the integration of new research findings in oncology. For this purpose, a semi-automatic mapping of ICD-O tuples to OncoTree codes was developed. The implementation uses a FHIR terminology server, pre-coordinated or post-coordinated SNOMED CT expressions, and subsumption testing. Various validations have been applied. The results were compared with reference data of scientific papers and manually evaluated by a senior pathologist, confirming the applicability of SNOMED CT in general and its post-coordinated expressions in particular as a viable intermediate mapping step. Resulting in an agreement of 84,00 % between the newly developed approach and the manual mapping, it becomes obvious that the present approach has the potential to be used in everyday medical practice.

Keywords. ICD-O, OncoTree, SNOMED CT, Ontoserver, HL7 FHIR, post-coordination, terminology server

1. Introduction

In a Molecular Tumor Board (MTB), an interdisciplinary team of physicians creates therapy recommendations for patients with oncological diseases beyond standard treatment options. In the MTB of the University Medical Center Schleswig-Holstein (UKSH), the software cBioPortal [1] shall be used to visualize molecular genetics and clinical data and to support decision-making. cBioPortal uses OncoTree, a hierarchically organized structure for the classification of currently 868 tumor types [2]. By considering a tumor's histology and localization, it can be matched to a node of the OncoTree.

In pathology reports, neoplasms are routinely coded using the ICD-O classification [3]. ICD-O differentiates between codes of two axes – topography and morphology – which are combined into a tuple. While a mapping from OncoTree to ICD-O is available

[1] Corresponding Author, Tessa Ohlsen, Institute of Medical Informatics, University of Lübeck, Ratzeburger Allee 160, 23564 Lübeck, Germany; E-mail: tessa.ohlsen@student.uni-luebeck.de

[2], the given associations are inherently unidirectional due to ICD-O being far more granular. Although ICD-O and OncoTree are both classifications for the oncology domain, individual class boundaries are not necessarily similar and thus class extensions only overlap partly. Due to this discrepancy, a simple translation from ICD-O to OncoTree is troublesome so that a more sophisticated approach is needed. For our project, we hypothesized that SNOMED CT may work as a feasible intermediate mapping step. To additionally minimize the workload and sources of error during the mapping process, a multi-step procedure was developed to map ICD-O tuples to OncoTree codes in a preferably (semi-)automatic way.

2. Methods

Four steps were developed to map the ICD-O tuples to OncoTree. SNOMED CT, as the most comprehensive terminology in medicine [4], was used as a purpose-agnostic intermediate representation to mediate between ICD-O and OncoTree. Figure 1 shows an overview of the individual mapping steps, which are explained in the following. All steps of the mapping were implemented using a local instance of the HL7 FHIR (Fast Healthcare Interoperability Resources)-based terminology server Ontoserver [5].

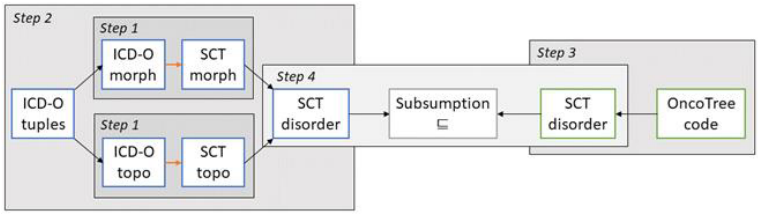

Figure 1. Sequence of mapping ICD-O tuples to the OncoTree with the four steps: (1) Automapping, (2) mapping of ICD-O tuples to SNOMED CT (SCT) disorder codes, (3) mapping of OncoTree codes to SNOMED CT disorder codes, and (4) subsumption testing between SCT-mapped ICD-O tuples and SCT-mapped OncoTree codes.

2.1. Automapping

In this preliminary mapping step, the two axes of ICD-O tuples (topography and morphology) are considered independently. All existing topography and morphology codes are provided by the World Health Organization (WHO), totaling up to 327 topography and 1090 morphology codes [3].

Each of the ICD-O topography and morphology codes shall be converted into SNOMED CT concepts. To achieve this, an Ontoserver-associated web application called Snapper[2] can be employed. Despite being primarily an editing tool for FHIR-based terminology resources, Snapper also offers an automapping feature, which has proven to be reliable and efficient when mapping to SNOMED CT [6]. So, the ICD-O codes are imported into Snapper and automatic mapping suggestions generated with appropriate settings: The target range for ICD-O topography and morphology codes can be limited to SNOMED CT concepts of the subhierarchies of body structures and morphologic abnormalities, respectively.

[2] https://ontoserver.csiro.au/snapper2/

Afterwards, a manual post-processing (choosing from suggestions or augmenting them via manual search) is completed by the PhD student. The resulting relations between ICD-O topography or morphological codes and the corresponding SNOMED CT concepts are stored in separate FHIR ConceptMaps [7] on the Ontoserver.

2.2. Mapping of ICD-O tuples to SNOMED CT disorder codes

Based on the previous step, the mapping of combined ICD-O tuples is performed. Here, the input dataset consists of 1800 ICD-O tuples used in tumor documentation at UKSH, Campus Lübeck since 2016. For each tuple, the two SNOMED CT codes corresponding to its topography and morphology can be used as the basis for the semi-automated detection of a pre-coordinated concept or the automated generation of a post-coordinated expression. In both cases, *64572001 | Disease |* is used as the central "focus" concept which is further refined according to the SNOMED CT Concept Model via the attributes *363698007 |Finding site|* and *116676008 |Associated morphology|* with the respective body structure and morphologic abnormality concepts. Like before, mapping results are collated into two separate ConceptMaps.

Pre-coordination

The pre-coordinated approach uses the SNOMED CT Expression Constraint Language (ECL) to find predefined concepts which fulfill the given expression. According to the basic structure described above, an example ECL expression is as follows:

< 64572001 | Disease | :
{ 363698007 | Finding site | = 39607008 | Lung structure | ,
116676008 | Associated morphology | = >! 35917007 | Adenocarcinoma | }

This expression queries all diseases which are found at the lungs with a morphology of adenocarcinoma or one of its direct parent concepts and would yield *707451005 |Primary adenocarcinoma of lung|* as a result.

An iterative algorithm was developed for retrieving a fitting pre-coordinated concept. In the first iteration, it considers only the exact attribute-value pairs as defined above. In subsequent iterations, further levels of parent concepts are considered for the topography and/or morphology concept, making the expression increasingly more general. The algorithm terminates as soon as at least one result is found or after a maximum of 14 iterations. If multiple results are returned, the best option is chosen interactively.

Post-coordination

The second approach makes use of SNOMED CT Postcoordinated Expressions (PCE) which allow for the flexible combination of multiple concepts into previously unrepresentable meanings. Thus, a granularity beyond the scope of predefined concepts is conceivable. While PCEs look similar to ECL expressions on a syntactic level, they represent valid SNOMED CT "concepts" instead of queries. So, each ICD-O tuple is mapped to a PCE constructed according to the partwise mapping results and the basic structure described above, e.g.:

64572001 | Disease | :
{ 363698007 | Finding site | = 39607008 | Lung structure | ,
116676008 | Associated morphology | = 35917007 | Adenocarcinoma | }

2.3. Mapping of OncoTree codes to SNOMED CT disorder codes

The 868 OncoTree codes are converted manually to SNOMED CT disorder codes by the PhD medical student. Here again, post-coordinated SNOMED CT expressions are built following the known structure and mapping results stored in a separate ConceptMap on the Ontoserver.

2.4. Subsumption testing and overall mapping

After mapping both the ICD-O tuples and OncoTree codes to SNOMED CT disorder codes, subsumption relations between both can be identified pairwise by querying the Ontoserver with the HL7 FHIR operation "$subsumes". Only the results "equivalent" and "subsumes" imply a relevant association from the respective ICD-O tuple to the OncoTree code which is then stored in yet another ConceptMap for the overall mapping.

3. Results

From the previously mentioned dataset, 105 of 1800 ICD-O tuples were identified as invalid and had to be excluded. Of the remaining 1695 ICD-O tuples used, 99.23 % could be successfully mapped to an OncoTree code using the pre-coordinated approach. With post-coordination, a mapping could be achieved for all input tuples. For 63.24 % of mapping relations, the selected target OncoTree codes are equivalent between both approaches with the code chosen via post-coordination being more specific and thus more precise otherwise.

The most frequent 100 ICD-O tuples already cover 63.00 % of all oncological diseases registered at UKSH, Campus Lübeck since 2016. A senior pathologist previously not involved in the process manually mapped this excerpt as reference for determining mapping accuracy. 84.00 % were found to be equivalent with the post-coordinated approach and 56.00 % using pre-coordination. For another validation, a previously published mapping by Thomas et al. [8] was used as reference data. 77.92 % of results based on the superior post-coordinated approach matched with the reference data. Otherwise, the reference data were 1.48 levels deeper in the OncoTree on average.

4. Discussion

By implementing a multi-step process, a semi-automated mapping from ICD-O tuples to OncoTree codes could be achieved successfully and with decent accuracy. SNOMED CT was found to be a workable solution to both bridge the gap between disparate classifications and to support automatization, especially attained by employing advanced features like ECL, post-coordination, and subsumption testing. Utilizing post-coordinated expressions also proved useful in achieving more precise mapping results by covering a broader scope as well as by preventing the loss of information inevitable when limited to the predefined combinations of pre-coordinated concepts.

A pre-requisite for implementing the described mapping approach was the availability of appropriate tooling. Here, the FHIR-based Ontoserver convincingly

supported the standardized access of terminology content and related operations, including the previously described specialized SNOMED CT features.

Nevertheless, validation revealed some inaccuracies in the mapping results which can be mainly attributed to three issues. Firstly, during validation, only a binary measure for equivalence was utilized. But, despite being not exactly the same, many results are still semantically similar. Secondly, OncoTree currently provides - with only 868 classes - significantly less detail compared to ICD-O and SNOMED CT, making the mapping inherently imprecise and non-reversible. Thirdly, further discrepancies between the involved terminologies hinder the mapping process. E.g., OncoTree's structure follows pragmatic considerations of everyday clinical practice which are sometimes incompatible with SNOMED CT's strictly logical polyhierarchy.

To mitigate some of these issues, further evaluations considering the semantic distance between divergent results and the specific influence of using SNOMED CT as an intermediate representation are in progress.

5. Conclusion

A largely automated mapping of ICD-O tuples to OncoTree codes could be implemented successfully by using SNOMED CT as an intermediate step. SNOMED CT, in combination with HL7 FHIR operations and a terminology server, enables a straightforward implementation. The approach using post-coordination outperformed the pre-coordination variant both in mapping coverage and accuracy. The results can easily be expanded to further ICD-O tuples and will be integrated into cBioPortal in the future.

Acknowledgements

This work was supported by the HiGHmed project within the Medical Informatics Initiative, funded by the Federal Ministry of Education and Research (grant number 01ZZ1802Z).

References

[1] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013 Apr 2;6(269):pl1.

[2] Kundra R, Zhang H, Sheridan R, Sirintrapun SJ, Wang A, Ochoa A, et al. OncoTree: A Cancer Classification System for Precision Oncology. JCO Clinical Cancer Informatics. 2021 Mar;(5):221–30.

[3] Staff WHO, Organization WH, Jack A, Percy C, Sobin L, Whelan S. International Classification of Diseases for Oncology: ICD-O. World Health Organization; 2000. 252 p.

[4] SNOMED - 5-Step Briefing [Internet]. SNOMED. [cited 2022 Jan 20]. Available from: https://www.snomed.org/snomed-ct/five-step-briefing

[5] Metke-Jimenez A, Steel J, Hansen D, Lawley M. Ontoserver: a syndicated terminology server. J Biomed Semantics. 2018 Sep 17;9(1):24.

[6] Drenkhahn C, Burmester S, Ballout S, Ulrich H, Wiedekopf J, Ingenerf J. Using FHIR terminology services-based tools for mapping of local microbiological pathogen terms to SNOMED CT at a German university hospital. In 2020. Available from: https://dvmd.de/wp-content/uploads/2020/11/A-148.pdf

[7] ConceptMap - FHIR v4.0.1 [Internet]. [cited 2022 Jan 21]. Available from: https://www.hl7.org/fhir/conceptmap.html

[8] Thomas S, Lichtenberg T, Dang K, Fitzsimons M, Grossman RL, Kundra R, et al. Linked Entity Attribute Pair (LEAP): A Harmonization Framework for Data Pooling. JCO Clinical Cancer Informatics. 2020 Nov 1;(4):691–9.