Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210126

Evaluating Suitability of SNOMED CT in Structured Searches for COVID-19 Studies

Carina Nina VORISEK[a,1], Sophie Anne Ines KLOPFENSTEIN[b], Julian SASS[a], Moritz LEHNE[a], Carsten Oliver SCHMIDT[c], and Sylvia THUN[a,d]
[a]*Berlin Institute of Health at Charité, Universitätsmedizin Berlin*
[b]*Institute of Medical Informatics, Charité Universitätsmedizin Berlin, Berlin*
[c]*University Medicine of Greifswald*
[d]*Hochschule Niederrhein - University of Applied Sciences, NFDI4Health Task Force COVID-19, Krefeld*

Abstract. Studies investigating the suitability of SNOMED CT in COVID-19 datasets are still scarce. The purpose of this study was to evaluate the suitability of SNOMED CT for structured searches of COVID-19 studies, using the German Corona Consensus Dataset (GECCO) as example. Suitability of the international standard SNOMED CT was measured with the scoring system ISO/TS 21564, and intercoder reliability of two independent mapping specialists was evaluated. The resulting analysis showed that the majority of data items had either a complete or partial equivalent in SNOMED CT (complete equivalent: 141 items; partial equivalent: 63 items; no equivalent: 1 item). Intercoder reliability was moderate, possibly due to non-establishment of mapping rules and high percentage (74%) of different but similar concepts among the 86 non-equal chosen concepts. The study shows that SNOMED CT can be utilized for COVID-19 cohort browsing. However, further studies investigating mapping rules and further international terminologies are necessary.

Keywords. Semantic interoperability, standardization, SNOMED CT, COVID-19

1. Introduction

The ongoing COVID-19 pandemic, caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), has triggered large numbers of health studies. Therefore, the NFDI4Health Task Force COVID-19 initiative, a German research project, was established to centralize clinical, epidemiological and public health studies on COVID-19 in Germany [1]. Interlinked complementary web-based platforms [2][3] provide study related information, exploration, visualization tools and storage of semantically annotated items. The NFDI4Health Task Force COVID-19 identified surveys and datasets for inclusion. To provide FAIR data - i.e., data that are findable, accessible, interoperable and reusable - data elements need to be interoperable [4]. This requires the use of established, standardized vocabularies, terminologies or ontologies such as SNOMED CT.

[1] Carina Nina Vorisek, Core Unit eHealth and Interoperability (CEI), Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Anna-Louisa-Karsch-Str. 2, 10178 Berlin, Germany; E-mail: carina-nina.vorisek@charite.de

As an example dataset to evaluate usage of SNOMED CT within the NFDI4Health Task Force COVID-19, we used the German Corona Consensus Dataset (GECCO), an interoperable pioneer dataset providing data items for standardized COVID-19 research [5]. GECCO was developed within the German National Research Network of University Medicine on COVID-19 ("Netzwerk Universitätsmedizin") and specifies information to be collected in COVID-19 studies. The GECCO core data set contains 81 elements combined with 281 response options, classified into 13 categories (Figure 1). In order to enhance semantic and syntactic interoperability of research data on SARS-CoV-2, GECCO uses Fast Healthcare Interoperability Resources (FHIR) to define machine-readable data formats for interoperable data exchange. In addition, GECCO dataset elements are mapped to international terminologies (SNOMED CT, LOINC, UCUM, ICD-10-GM and ATC) to ensure semantic interoperability. We performed a mapping of GECCO using SNOMED CT. SNOMED CT concepts enable stable mechanisms capturing information and supporting data integration. However, the understanding of terms of a terminology may differ, implementing challenges to both developers and users.

This study evaluates intercoder reliability and suitability of mapping SNOMED CT to GECCO. Results are, for example, of relevance for the development of standardized filtering and browsing options in COVID-19 study search portals, such as the ones developed within the NFDI4Health Task Force COVID-19.

2. Methods

2.1. Mapping procedure

Within the GECCO dataset, we identified 205 data items (response options representing simple yes/no/unknown responses were considered as straightforward and excluded from analysis). Mapping was conducted independently by two coders (CV, SK), both medical doctors with broad experience in the area of medical terminologies. No mapping rules were established prior to the mapping process. Mapping was conducted with the SNOMED CT browser (Release: International Edition 2020-07-31).

Concepts behind each of the 205 data items were analyzed briefly and mapped manually to SNOMED CT. Both coders (CV, SK) mapped each data element to SNOMED CT concepts and then decided which concept was suitable for each individual data element. A joint mapping version was established based on consensual validation by both coders. Mapping results of the joint mapping version were rated and classified using ISO/TS 21564:2019 "Health Informatics — Terminology resource map quality measures (MapQual)" to measure the degree of equivalence between the dataset and target concept of SNOMED CT (Table 1) [6].

2.2. Intercoder Reliability

Mapping results between the two coders were compared among each other and with the original SNOMED CT mapping of the identified GECCO data items. Percentage share of SNOMED CT concepts to the GECCO dataset equally chosen by the coders and nominal Krippendorff's α was used as a measure of intercoder agreement. Different mapping choices were analyzed and divided into similar and non-similar concepts among the two coders.

3. Results

3.1. ISO/TS 21564 Equivalence Assessment Score

The joint mapping version of the GECCO dataset with SNOMED CT achieved a high average ISO/TS 21564 equivalence assessment score of 0.52 (Table 1). Overall, 69% of data items had equivalent meaning in the chosen SNOMED CT concept. All but one data item had a match in the used SNOMED CT system. ISO equivalence assessment by data category is presented in Figure 1. Fifteen (7%) data items had multiple target concepts (e.g., terms describing a conjunction of several diseases), and 189 (92%) data items had a unique respective SNOMED CT concept.

Table 1. ISO/TS 21564 Equivalence Assessment Score: ~0.52

Rating	Description	Number of Concepts (%)	Examples	
			Original GECCO Data Item	SNOMED CT Concept
0	Equivalent meaning	141 (69)	Dementia	Dementia (disorder)
1	Source is wholly included in target	32 (16)	Laboratory Test: Cardiac Troponin	Troponin measurement (procedure)
2	Source is partially included in target	23 (11)	Respiratory Outcome	Dependence on ventilator (finding)
3	Source is mapped however there were many options. Source map is the best comparison rather than an actual correspondence	8 (4)	Does the patient suffer from at least one rheumatological/ immunological disease?	Rheumatism (disorder) / Disorder of immune function (disorder)
4	No mapping possible	1 (1)	Symptoms: Other	/

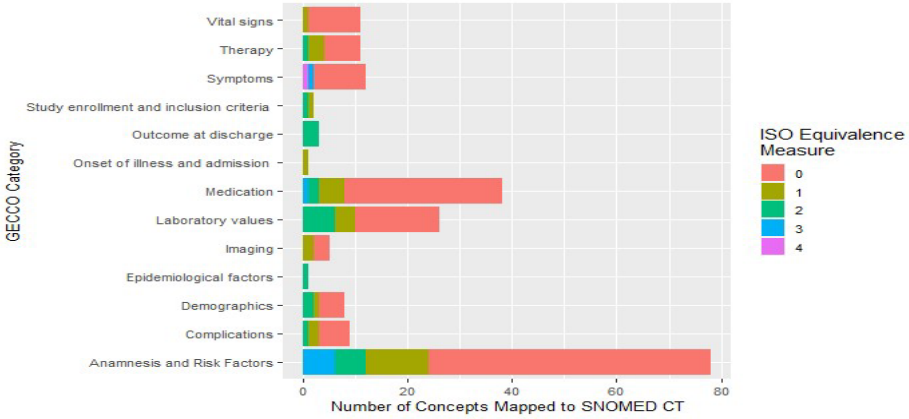

Figure 1. ISO/TS 21564 Equivalence Assessment Score by GECCO categories (0: equivalent meaning; 1: source is wholly included in target; 2: source is partially included in target; 3: source is mapped however there were many options; 4: no mapping possible).

3.2. Intercoder Reliability

Results on intercoder reliability are shown in Table 2. In the original GECCO data set, 61 (30%) of the items were not mapped to SNOMED CT resulting in 144 items which were compared to the two coders (Table 2).

Table 2. Intercoder Reliability (percentage share of SNOMED CT concepts to the evaluated GECCO dataset equally chosen by the coders) and results for nominal Krippendorff's α

		N (%)	Krippendorff's α
		N = 144	
GECCO - CV			
	No Match	73 (51)	0.49
	Match	71 (49)	
GECCO - SK			
	No Match	70 (49)	0.50
	Match	74 (51)	
		N = 205	
CV - SK			
	No Map	1 (0)	
	No Match	86 (42)	0.57
	Match	118 (58)	

Analyzing the 92 (45%) non-equal mapping concepts of the two coders, we found that 64 (74%) concepts were different but related or similar concepts from SNOMED CT (Table3).

Table 3. Examples of different but similar SNOMED concepts

Original GECCO Data Item	SNOMED CT concept chosen by coder CV	SNOMED CT concept chosen by Coder SK
Does the patient have a history of being an organ transplant recipient? Skin	Transplanted skin present (finding)	History of skin recipient (situation)
Type of ventilation: Non-invasive ventilation	Dependence on non-invasive ventilation (finding)	Noninvasive ventilation (procedure)
Symptoms: Vomiting	Vomiting (disorder)	Finding of vomiting (finding)

4. Discussion and Conclusion

In this article, we showed that SNOMED CT concepts capture typical data elements relevant to COVID-19 studies, exemplified by the GECCO dataset, very well. The mapping results reached a high ISO/TS 21564 equivalence assessment score. Sass et al. standardized Germany's Electronic Disease Management Program for bronchial asthma using standards such as LOINC, UCUM as well as SNOMED CT and found comparable high mapping results [7]. SNOMED CT could therefore be used to enable detailed structured searches, for example in study portals like the one developed by the NFDI4Health Task Force COVID-19 and beyond. However, not all data elements could be perfectly matched to SNOMED CT, and for improved accuracy, a hybrid approach using several terminologies might be preferable.

Intercoder reliability between the two coders in our study was moderate, however, the majority of non-equal matches were identified as similar concepts. An

elevated count of discrepancies may be due to the fact that no rules were determined prior to the mapping as improvement of intercoder results after establishment of mapping rules has been previously reported [8].

In the future, implementation of mapping rules is necessary to improve mapping quality and intercoder reliability. We plan to further facilitate browsing of COVID-19 research information within the NFDI4Health Task Force COVID-19 project and will continue to evaluate the possibility of a faceted search of study information using SNOMED CT.

Acknowledgements

This work was done as part of the NFDI4Health Task Force COVID-19 (www.nfdi4health.de). We gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project Number 451265285, PI 345/17-1/SCHM 2744/9-1.

References

[1] Task Force COVID-19 - NFDI4Health. Available from: https://www.nfdi4health.de/de/task-force-covid-19. Accessed Feb 20201.
[2] Wolstencroft K, Owen S, Krebs O, Nguyen Q, Stanford NJ, Golebiewski M, et al. SEEK: A systems biology data and model management platform. BMC Syst Biol. 2015 Jul 11;9(1).
[3] Doiron D, Marcon Y, Fortier I, Burton P, Ferretti V. Software application profile: Opal and mica: Open-source software solutions for epidemiological data management, harmonization and dissemination. Int J Epidemiol. 2017 Oct 1;46(5):1372–8.
[4] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3.
[5] Sass J, Bartschke A, Lehne M, Essenwanger A, Rinaldi E, Rudolph S, et al. The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond. BMC Med Inform Decis Mak. 2020 Dec 1;20(1):341.
[6] International Organization for Standardization, ISO/TS 21564: Health Informatics – Terminology resource map quality measures (MapQual).
[7] Sass J, Essenwanger A, Luijten S, et al. Standardizing Germany's electronic disease management program for bronchial asthma. In: Studies in Health Technology and Informatics. IOS Press, 2019: 81–85.
[8] Vikström A, Skånér Y, Strender LE, Nilsson GH. Mapping the categories of the Swedish primary health care version of ICD-10 to SNOMED CT concepts: Rule development and intercoder reliability in a mapping trial. BMC Med Inform Decis Mak. 2007;7:9.