Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240687

'SNOMEDizing' Questionnaires for Standardizing Stroke Registry Data

Andrea RIEDEL[a], Stefan SCHULZ[b] and Catalina MARTÍNEZ COSTA[c,1]
[a] *Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany*
[b] *Institute for Medical Informatics, Statistics and Documentation, Medizinische Universität Graz, Graz, Austria*
[c] *Department of Informatics and Systems, Universidad de Murcia, CEIR Campus Mare Nostrum, IMIB-Arrixaca, Murcia, Spain*
ORCiD ID: AR https://orcid.org/0009-0006-9389-9482, SS https://orcid.org/0000-0001-7222-3287, CMC https://orcid.org/0000-0003-1857-1744

Abstract. International interoperability of healthcare and research data requires a commitment to standards. To this end, SNOMED CT was evaluated for representing questionnaire items of the European Registry of Stroke Care Quality using a complex annotation protocol. The agreement between validators and annotators was 72.4%. At least 64% of the information could be represented by using SNOMED CT only, including complex post-coordinations. 9% of the information would require an information model, and 14% the addition of new content to SNOMED CT. Next steps will be the creation of an annotation guideline for questionnaires, a specific reference set, and the combination of both with an information model such as HL7 FHIR.

Keywords. Stroke, SNOMED CT, Information model, Interoperability

1. Introduction

Stroke is a global health problem with significant economic impact. To improve stroke research and management, interoperable data of high quality are mandatory. This became evident with RES-Q, a large international stroke quality register [1], launched in 2016, with more than 2,500 hospitals and 790,000 patients in 92 countries. Using a single data collection form via a questionnaire, primary and comprehensive stroke centers collect data on acute and post-acute stroke treatment.

The Horizon Europe project RES-Q+ aims at semantically standardizing this questionnaire using international interoperability standards (e.g., SNOMED CT (SCT) [2], HL7 FHIR [2] or OMOP CDM [3]). A semantic interoperability layer is built upon the Semantic Web standards RDF and OWL. At its core, the RES-Q+ ontology is intended to cover all data items in RES-Q. This work represents initial steps including the identification of existing ontologies suited for RES-Q data, the semantic annotation of the questionnaire to SCT and its coverage analysis.

[1] Corresponding Author: Catalina Martínez-Costa; E-mail: cmartinezcosta@um.es.

2. Methods

Data collected with a unique questionnaire encompass hospitalization, treatment and post-treatment of stroke patients. It consists of around 124 single and multiple-choice questions. Free-text elements were strongly limited to a few date/time and numeric measurements in a predefined format. To represent RES-Q registry data, the RES-Q+ ontology is built to facilitate data analysis and interoperability with other registries. SCT is the ontology selected due to its wide coverage of healthcare. It is embedded in a framework that represents information in a contextualized way, such as adding temporal aspects, numeric values, or provenance [4]. We here present an analysis of the RES-Q questionnaire annotation with SCT, resulting in a knowledge graph with all registry data.

Encoding detailed clinical information with SCT requires both medical domain expertise and terminology skills [5]. Usually, medical experts receive short training based on an annotation guideline. However, when this task becomes complex, particularly when annotations need to be logically connected by post-coordination [6], more intensive preparation and a continuous interdisciplinary exchange are mandatory. To balance annotating multifaceted phrases and a straightforward introduction to the task, the complex RES-Q+ annotation followed successive, incremental steps involving experts on both domains, stroke and SCT. Two SCT experts (*AR*, *CMC*), although familiar with stroke, provided the first annotation of the English-language questionnaire using the international SCT edition via the SCT web browser. Additionally, they added an annotation status (open/in progress/finalized) and correlation [7] to track the process by necessary metadata. Based on their experience, the annotators developed a simple annotation guideline and validation schema with three levels: (i) green = '*agreement with annotation of the SCT expert*', (ii) yellow = '*..., but like to add an improvement/ suggestion as a comment*' or (iii) red = '*disagreement with annotation of the SCT expert and adding of own recommendation or reference to co-validators comment*'. In a second step, six stroke experts reviewed the proposed annotation from a medical point of view after receiving a short introduction to SCT and the validation task. Then, each of them validated a quarter of the questionnaire regarding the correct representation of the clinical meaning by the pre-chosen SCT concept. As a third step, an expert on both domains (*SS*), reviewed all annotations. Where disagreements could not be settled, input from another medical stroke expert was provided. Finally, based on the annotation and validation process, the SCT experts (*AR*, *CMC*) analyzed how many questions and answers (Q/As) could be annotated with SCT and the necessary changes due to validators' feedback.

3. Results

To simplify the medical review, '*Yes/no*' answers indicating presence/absence of a condition or procedure status were left out, because this can be represented by post-coordination guided by the SCT context model. Yet we represented such contextual cues within the RES-Q+ ontology, to facilitate future representations using information model (IM) standards such as HL7 FHIR or OMOP CDM. For 417 Q/As and headings, an agreement of 72.4% between the SCT annotators and both medical validators could be shown. In 14.4% only one medical validator agreed with the annotators, and in 13.2% both validators offered the same or different alternatives. 303 different SCT concepts were used via pre-coordination or combined for post-coordination. Their annotation could be divided into the following variants (Table 1).

Table 1. SCT annotation variations and their percentages.

SCT annotation		Rate	Question	Answer
1	**Pre-coord.**	**35%**	Brain imaging type	Non-Contrast CT
			396205005 | Computed tomography of brain without radiopaque contrast (procedure) |	
2	**Post-coord.**	**29%**		
2.1	Simple post-coord. indicating AND/OR	3%	Stroke mimics final diagnosis	Electrolyte or metabolic imbalance
			105593004 | Electrolyte imbalance (disorder) | OR *75934005 | Metabolic disease (disorder) |*	
2.2	More complex post-coord.	26%	Antidote anticoagulant given	Yes
			67329000 |Administration of antidote (procedure)| : { 260686004 |Method (attribute)| = 129445006 |Administration - action (qualifier value)|, 363701004 |Direct substance (attribute)| = 443017007 | Coagulation related substance (substance) | }	
3	**impossible**	**25%**		
3.1	but belongs to IM	9%	Imaging done at what time?	*date time value*
			COVID positive in last 6 months	Yes/No
3.2	no support of intended post-coord.	2%	Age at stroke onset	*Integer value*
			The patient hospitalized in (day 1)	(Department) with monitored bed with telemetry
3.3	concept missing	14%	ICH score in intracerebral hemorrhage	Integer value (0-6)
			Remote bleeding in the brain	Yes/No
4	**(An)other**	**11%**	Any post-stroke complication?	Other

Most common was a direct annotation to SCT via pre-coordination [8] (1 in Table 1/ 35%). Often a pre-coordinated SCT concept alone did not represent data well enough, because context was needed. When a pre-coordinated concept existed, we ignored how and when the information had been acquired (e.g., at hospital admission, or during post-acute care). This usually corresponds to questionnaire headings and was represented within the IM. E.g., that the first systolic blood pressure measurement was taken on admission, this procedure was part of the hospital admission procedure and could only be expressed using an IM. Another case was the representation of brain infarcts seen on an image, with many possible image types (CT, MR DWI/FLAIR, etc.). Again, the linkage of the old brain infarct should be related to an image type within the IM.

29% of 'Q/A' items require SCT post-coordination (2 in Table 1), being 3% simple 'AND/OR' combinations of pre-coordinated concepts (2.1) and 26% complex expressions (2.2) with more than one domain-attribute-range combination, typically including procedures and conditions at body locations, drug administration, and prescription. Many have temporally and causally dependencies, such as '*dissection of artery as a thrombectomy complication*', represented as *710864009 |Dissection of artery (disorder)|: 255234002 |After (attribute)| = 43810009 |Removal of thrombus (procedure)|*. Within (2.2) we also found concepts related by '*associated with*' or '*due to*', although temporal information is typically represented within an IM. While the SCT context model provides some temporally qualified concepts (e.g., '*previous myocardial infarction'*), when expressing e.g., '*Covid positive in last 6 months*' or '*Patient hospitalized (day 1)*', the IM is needed.

25% of the cases were impossible to represent with SCT only (3 in Table 1). 9% of the annotation included information that required the IM (3.1), mostly for reasons of actions and numeric temporal information. Temporal information was often complex

(3.2), e.g., '*Age at stroke onset*', '*3-months Modified Rankin Scale (mRS)*' or '*TIA symptoms duration*'. Such information is usually encoded by SCT Observables, which are, however, primitive concept, and raise the question about enabling post-coordination for observables, too. These cases (3.2/ 2%) fall, together with more complex ones such as '*(Department/Unit) with monitored bed with telemetry*', in which the device needs to be linked to bed and bed to hospital department. In 14% (3.3), SCT annotation was impossible for lack of coverage, such as '*Intracerebral Hemorrhage Score (ICH)*'or '*ASSIST swallowing test*'. Moreover, codes for some substances did not exist, as well as for complex expressions, e.g., '*Remote bleeding in the brain*' or '*Out of time window*'.

Finally, residual expressions with '*another*', '*unspecified*', and '*elsewhere*' were common in answers on conditions (4 in Table 1/ 11%). For the sake of simplicity and to avoid negation, e.g., '*Other Anticoagulant*' was interpreted as '*Any anticoagulant*' (i.e., *372862008 | Anticoagulant (substance) |*). However, this was not valid in cases such as '*Other'* as reason for post-stroke complications, given the wide range of possible concepts.

4. Discussion and Conclusions

With a complex annotation protocol and the participation of stroke experts and medical terminologists, we showed that at least 64% of the questionnaire's information could be represented using SCT only. 9% of the information required an IM, and in 14% it was necessary to add new concepts to SCT. Despite the absence of guidelines and reference sets for stroke questionnaires, we were able to annotate more than 75% of the extensive RES-Q questionnaire. Future work will benefit from specific guidelines for questionnaire content [9,10]. In addition, user feedback to SNOMED International is necessary and more guidance to use SCT codes in the context of an IM. Both were demonstrated by the amount of missing SCT content and content that typically belongs to an IM. A predefined stroke reference set would facilitate future refinements, as well as the representation of contextual aspects using HL7. Another challenge is given by the annotation of narrative information [11]. The annotation of the RES-Q questionnaire with SCT highlighted main challenges of encoding clinical data which are discussed in the following:

The capture of registry data as in RES-Q+ creates awareness that workflows need to be improved and interface between SCT and clinicians be strengthened [5,9]. Such scenarios must address the trade-off between limited implementation resources and the need to exchange knowledge about terminology, in particular targeting medical experts.

Limited terminology precision, demonstrated by limited SCT coverage [12], is a severe limitation. It includes codes for categories such as antidiabetics, as well as for specific tests, e.g., swallowing tests. Moreover, monthly SCT updates complicate its use in short project-related annotation tasks [5]. Just as challenging is post-coordination, because some necessary post-coordination patterns violate SCT's domain/range constraints and would therefore by better addressed by the IM.

Language peculiarities, e.g., '*Yes/no*' answers are typical for questionnaires, and multiple-choice questions often include the items '*unspecified*' and '*others*'. The correct interpretation of '*other*' as complement would require logical negation, which is not included in OWL-EL, the logic of SCT. Negation with exclusions of the already mentioned answer options should be implemented with the IM.

Missing precision of the linguistic design requires that questionnaire items be formulated in a clear language. Double-checking content by experts is a requirement for

high accuracy [5]. This applies particularly to translations, which must not alter the meaning [12]. To this end, term translations need alignment with term translations of localized SCT versions. Abbreviations, geographic, and institutional term preferences should be avoided. Questionnaire authors must emphasize broad understandability. Standardized multilingual terminologies such as SCT with their preferred terms in several languages help focus on the broad understandability of the terminology used.

Declarations

Funding: Fellowship within IFI program of German Academic Exchange Service (DAAD) and funded by Federal Ministry of Education and Research (BMBF) (AR). Grant RYC2020-030190-I, Horizon Europe HORIZON-HLTH-2021-TOOL-06-03 under Grant 101057603, and Horizon-HLTH-2022-Tool-12-Two-Stage under Grant 101080875 (CMC); **Authors' contributions:** Annotation and evaluation of validation (AR, CMC), technical validation (SS) - supervision (CMC) - original manuscript (AR) - review and editing (CMC, SS) - reading and approval of final manuscript (AR, CMC, SS). The present work was performed in (partial) fulfillment of the requirements for obtaining the degree 'Dr. rer. biol. hum.' from the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (AR); **Acknowledgements:** We want to thank all participants at the RES-Q+ project meeting (27.10.2024) for their interest, especially R. Mikulik, B. Casaru, E. Terescoasa, I. Antova, T. Queen, D. Taskov, and L. Palaiodimou; **Competing interests:** None declared.

References

[1] The stroke care quality registry. Res-Q. [cited 2024 May 21]. Available from: Stroke care quality registry | RES-Q.

[2] Benson T, Grieve G. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR. Cham: Springer International Publishing; 2016. doi:10.1007/978-3-319-30370-3.

[3] OMOP Common Data Model. [cited 2024 May 21]. Available from: https://ohdsi.github.io/CommonDataModel/.

[4] Abad-Navarro F, Martínez-Costa C. A knowledge graph-based data harmonization framework for secondary data reuse. Comput Methods Programs Biomed. 2024 Jan;243:107918. doi: 10.1016/j.cmpb.2023.107918. Epub 2023 Nov 10. PMID: 37981455.

[5] A. Riedel, N. Deppenwiese, L. Thiele, H.-U. Prokosch, A. Herzog, Development of a SNOMED CT Mapping Process and Tool at a Data Integration Centre - Lessons Learned, (Under Review for GMDS 2024).

[6] IHTSDO. SNOMED CT Glossary: Postcoordinated expression. [cited 2024 May 21]. Available from: SNOMED CT Glossary - SNOMED Confluence.

[7] IHTSDO. SNOMED CT Browser: Map source to map target correlation (foundation metadata concept). [cited 2024 May 21]. Available from: SNOMED CT - Map source to map target correlation (foundation metadata concept).

[8] IHTSDO. SNOMED CT Glossary: Precoordinated expression. [cited 2024 May 21]. Available from: SNOMED CT Glossary - SNOMED Confluence.

[9] Rossander A, Lindsköld L, Ranerup A, Karlsson D. A State-of-the Art Review of SNOMED CT Terminology Binding and Recommendations for Practice and Research. Methods Inf Med. 2021 Dec;60(S 02):e76-e88. doi: 10.1055/s-0041-1735167. Epub 2021 Sep 28. PMID: 34583415; PMCID: PMC8714300.

[10] Vorisek CN, Klopfenstein SAI, Sass J, Lehne M, Schmidt CO, Thun S. Evaluating Suitability of SNOMED CT in Structured Searches for COVID-19 Studies. Stud Health Technol Inform. 2021 May 27;281:88-92. doi: 10.3233/SHTI210126. PMID: 34042711.

[11] Schulz S, Miñarro Giménez JA, Karlsson D, Rosenbeck Gøeg K, Markó K, Van Bruwaene B, Cornet R, Jaulent M-C, Hämäläinen P, Chronaki C, Stroetmann V, Kalra D. ASSESS CT. Deliverable. D2.4 - Use of terminologies for representing structured and unstructured clinical content. [cited 2024 May 21]. Available from: assess_ct_ga_643818_d2.4.pdf (assess-ct.eu).

[12] van der Kooij J, Goossen WT, Goossen-Baremans AT, de Jong-Fintelman M, van Beek L. Using SNOMED CT codes for coding information in electronic health records for stroke patients. Stud Health Technol Inform. 2006;124:815-23. PMID: 17108614.