Intelligent Health Systems – From Technology to Data and Knowledge
E. Andrikopoulou et al. (Eds.)

doi:10.3233/SHTI250440

Semantic Representation of Medical Data Collection Forms Using Standards

Stefan SCHULZ[a, 1] and Catalina MARTÍNEZ COSTA[b]
[a] *Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria*
[b] *Department of Informatics and Systems, University of Murcia, CEIR Campus Mare Nostrum, IMIB-Arrixaca, Murcia, Spain*
ORCiD IDs: 0000-0001-7222-3287, 0000-0003-1857-1744

Abstract. Data collection forms and questionnaires are widely used to capture patient data, but their development often receives insufficient attention. The RES-Q+ project aims to semantically standardise a widely used international stroke registry data collection form using interoperability standards such as SNOMED CT and HL7 FHIR, pursuing a canonical, knowledge-graph-based representation as an ultimate goal. We analysed patterns in this RES-Q data collection form to derive general principles for improving the design and semantic alignment of medical forms. We studied the way how residuals such as "none", "other" and "unknown" can be modelled using SNOMED CT codes bound to FHIR resources.

Keywords. Stroke, SNOMED CT, FHIR, Terminology binding

1. Introduction

Clinical data collection forms (DCFs) provide structured information such as required by disease registries and retrospective studies [1]. However, the design of DCFs often lacks methodological rigour. They are typically developed in a bottom-up manner by clinicians and researchers, who are mostly not experts in DCF design, and who intuitively take the meaning of domain terms and the context in which they are used for granted. Often, DCFs are questionnaire-like, sometimes they resemble checklists. DCFs are typically related to a clinical topic, targeted to specified users, and consist of a series of elements, often in a question-answer (Q/A) style. The question part of a DCF element is followed by a variable number of single or multiple-choice answers, or alternatively to fields to be filled with integers, decimals, or date/time expressions. Whereas fully structured DCFs do not have free-text entry fields, this is allowed in semi-structured DCFs. Q/A elements are usually processed by the user in a linear order, but additional rules may direct users to sub-forms or allow them skipping part of the DCF.
This paper focuses on the RES-Q registry [2], a stroke quality database covering over 800,000 stroke centre patients from 92 countries. Its DCF collects data on acute and post-acute care (cf. Fig. 1). The EU-funded RES-Q+ project [3] has set off to semantically align this form with the international standards SNOMED CT and HL7 FHIR [4]. Semantic standardization of DCFs facilitates analysis and re-use of clinical data across

[1] Corresponding Author: Stefan Schulz; E-mail: stefan.schulz@medunigraz.at.

healthcare settings, languages and clinical databases, including content with different degrees of structures, from structured content of EHR systems and registries to semantic extracts from clinical narratives. To come closer to this goal, RES-Q+ has leveraged knowledge graphs as target representations based on RDF and OWL [5].

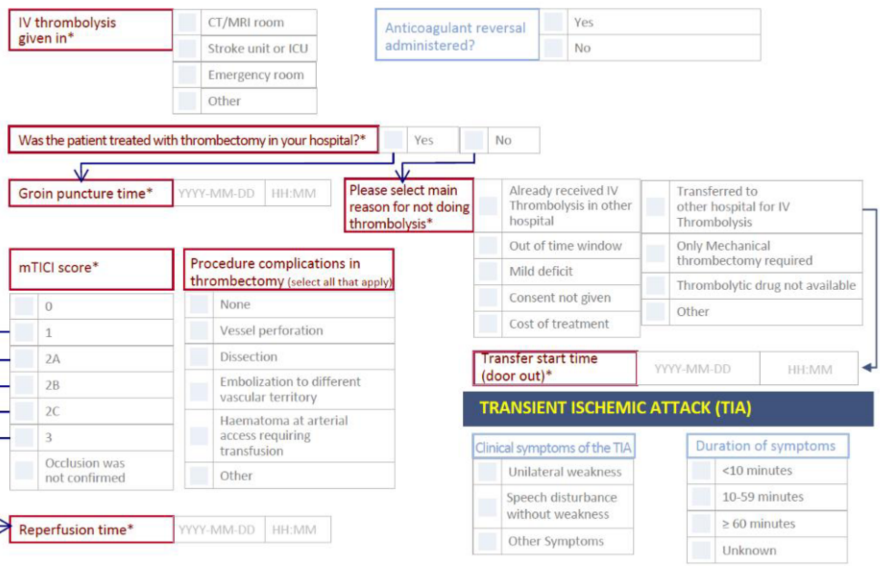

Figure 1. Part of the RES-Q clinical data acquisition form

2. Methodology

As described in [6], we identified general principles for the semantic specification of clinical DCFs. Our approach is based on the premise that integrating a detailed clinical ontology such as SNOMED CT with a set of templates as provided by FHIR can accurately represent the stroke information captured by the RES-Q DCF. This ensures that the resulting semantically rich data is compatible with similar data gathered elsewhere on the same clinical subject. We follow earlier work in the field of Applied Ontology, which makes a strict distinction between information entities (as typically represented by instances of FHIR resources) and clinical entities (as represented by SNOMED CT classes and instances thereof) [7].
Based on these principles, a qualitative analysis of typical examples is done with a focus on their semantic structure and the constituting pieces of meaning.

3. Results

The analysis of the RES-Q DCF revealed 88 question/answer (Q/A) elements, of which 57 were marked as mandatory. 21 dependencies were found between questions, which allowed multiple paths through the DCF, according to the answers selected (see arrows in Fig. 1). Table 1 presents the distribution of question types and subtypes according to special answer types. The subtypes were classified according to the occurrence of residuals, *viz.* answer items that refer to preceding parts of the Q/A, introducing negation or stating missing information.

Table 1. Distribution of RES-Q question types (n = 88)

Question/Answer (Q/A) Type	Subtype by residual	Freq Sub Type	Freq Type
Date / Time			11
Integer			13
	+ none	5	
Multiple choice (3-20 answers)			17
	+ none	2	
	+ none / other	4	
	+ none / other / unknown	2	
	+ other	5	
Single choice (2-9 answers)			34
	+ none	3	
	+ none / unknown	2	
	+ other	6	
	+ unknown	2	
String	(identifier, hospital names)		2
Yes / No			11
Total (Freq type)			88

Our analysis highlights the two complex Q/A types, *viz. Multiple Choice* and *Single Choice*. The *Yes/No* type is not set apart because it can be seen as a minimal form of *Single Choice*. The *String* type was instantiated only twice. Answer options, to be correctly analysed, often require contextual information conveyed by the DCF's Question part, as well as by the explicit and implicit characterisation of the entire DCF, which often requires an in-depth understanding of the domain and its context of use.

Commonalities: The FHIR resources *Condition*, *Procedure, Observation* and *Medication* were found sufficient to describe all context around the SNOMED CT codes bound to them. Similar DCFs may also require additional FHIR resources such as *FamilyMemberHistory* or *Immunization*. All resource instances can be assumed to be linked to the ID of the related treatment episode, from which patient and institution details are available via *subject* and *encounter* attributes, respectively.

Specific Answers: Each specific answer is related to a FHIR resource, which is instantiated whenever the answer is selected. The default assumption is that the related SNOMED CT code or expression refers to some instance in reality, i.e. it is contextualised as *confirmed present* unless otherwise stated. The attribute *code* represents the binding to a SNOMED CT concept (or, possibly a post-coordinated expression [7]) that characterizes the clinical concept the answer is about. E.g., "unilateral weakness" (under "TRANSIENT ISCHEMIC ATTACK – Clinical symptoms of the TIA" in Fig. 1) must be put in this context as a cause of this condition. This is accomplished by a post-coordinated SNOMED expression, as in the example below. In this example and the following ones, we employ a **Subject**; **Predicate**; **Object** triple notation. The **Subject** is given by the instance of the FHIR resource, which is instantiated once the answer item is selected. **Predicate** corresponds to the obligatory FHIR attribute **`code`**, with the **Object** filled by a SNOMED CT code or a post-coordinated expression employing SNOMED CT's compositional grammar. Additional triples refine the same FHIR resource instance by contextual information, e.g. employing the predicates **`verificationStatus`** or **`clinicalStatus`**. Nesting is possible; we here use variables (**`%1`**, **`%2`**) to maintain the triple structure:

```
fhir-condition#4z6pj; code; "20022000|Hemiparesis: 42752001|Due
to = 266257000|Transient ischemic attack"
```

Furthermore, information on factuality and temporality can be added:

```
fhir-condition#4z6pj; verificationStatus; "410605003|Confirmed
present"
fhir-condition#4z6pj; recordedDate; "2024-10-07T10:00:00Z"
```

Finally. the same condition instance is referred to regarding the duration of the symptoms, as an answer to "TRANSIENT ISCHEMIC ATTACK – Duration of Symptoms ":

```
fhir-condition#4z6pj; onsetRangeLow;  %1
fhir-condition#4z6pj; onsetRangeHigh; %2
%1; value; 10.0
%1; unit; "1156209001|Minute"
%2; value; 59.0
%2; unit; "1156209001|Minute"
```

Residuals: Residuals ("None", "Other", "Unknown") are frequent in the RES-Q form. Each of these three has a very distinct meaning: whereas "None" expresses negation (referring to some clinical state of affairs exposed in the Question part), "Other" excludes all sibling answer items, and "Unknown" specifies that there is no knowledge about that state of affairs at the moment the DCF is filled. "None" is treated like "No" in a *Yes-No* question, in the sense that the concept referred to in the question part is negated, e.g.

```
fhir-condition#4z6pj; code; "20022000|Hemiparesis"
fhir-condition#4z6pj; verificationStatus; "410516002|Known
absent"
```

"Other", if correctly interpreted, instantiates the concept referred to in the question part in a *positive* context, but all concepts related to the answer items (typically the siblings or the "Other" entries in the form) in a *negative* context. In the Fig. 1 example "Clinical symptoms of the TIA", the option "Other Symptoms" is modelled as follows:

```
fhir-condition#u11i8; code; "20022000|Hemiparesis"
fhir-condition#u11i8; verificationStatus;"410516002|Known absent"
fhir-condition#4yys4; code; "29164008|Disturbance in speech"
fhir-condition#4yys4; verificationStatus;"410516002|Known absent"
fhir-condition#ozzt5; code; "308921004|Neurological symptom:
42752001|Due to| = 266257000|Transient ischemic attack"
fhir-condition#ozzt5; verificationStatus; "410605003 |Confirmed
present|"
```

However, the scope of "Other" may go beyond this standard pattern. One case is "Stroke mimics" (everything that is not a stroke). Other examples are "Another hospital" and "Embolization to different vascular territory". A correct interpretation is therefore complex and requires the understanding of the domain knowledge, including knowledge about specific workflows, which is not explicitly stated in the form.

"Unknown", finally sets the verification status to the corresponding value. For instance, the answer item "unknown" under "Clinical history" could then be represented as follows:

```
Fhir:condition#45r62; code; "404684003|Clinical finding"
Fhir:condition#45r62; verificationStatus; "261665006|Unknown"
Fhir:condition#45r62; clinicalStatus; "410513005|In the past"
```

4. Discussion and Conclusions

We exploited the power of FHIR resources together with SNOMED CT codes and expressions bound to them, in order to semantically represent the content of clinical data collection forms (DCFs). The goal is to provide data with a comprehensive ontology-based grounding for further analysis and processing[2]. At least for our example we were able to show that FHIR allows for precise interoperable representations of patient-level meaning. Particularly, we demonstrated how the complex semantics of "other" could be expressed. According to the list of answer items, however, the representation of "other" items may become profuse, because the proposed solution entails an explicit exclusion statement for each sibling. Our approach is limited by the scope of the two standards, which does not offer solutions for disjunctions ("Carotid endarterectomy performed or stenting"), relative time references ("2 weeks after stroke"), and conditionals ("Was insulin administered for the first elevated glucose? (only if the level is >7mmol/L or 126 mg/dL)"). An unresolved issue is furthermore the overlap between SNOMED CT and FHIR [9]. Our strategy is to defer all epistemic, temporal and provenance information to FHIR and to avoid the use of the SNOMED Situation Hierarchy, e.g. 417662000 |History of clinical finding in subject (situation)|.

In conclusion, we recommend that frequently used clinical data collection forms be semantically standardised with the highest degree of precision. While this approach demands both expertise and time, it ensures the interoperability of data for potentially millions of patients, especially in large international initiatives like RES-Q.

Declarations

Funding: Grant RYC2020-030190-I funded by MCIN/AEI/10.13039/5011000110033/ and "FEDER: Una manera de hacer Europa", by the "European Union", Horizon Europe HORIZON-HLTH-2021-TOOL-06-03 under Grant 101057603, and Horizon-HLTH-2022-Tool-12-Two-Stage under Grant 101080875. **Competing interests:** None declared.

References

[1] Richesson RL, Nadkarni P. Data standards for clinical research data collection forms: current status and challenges. J Am Med Inform Assoc. 2011 May 1;18(3):341–346.
[2] The stroke care quality registry. Res-Q. https://www.stroke.qualityregistry.org/ [accessed 2025-01-06].
[3] RES-Q+ https://www.resqplus.eu [accessed 2025-01-06].
[4] 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.
[5] 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.
[6] Riedel A, Schulz S, Martínez Costa C. 'SNOMEDizing' Questionnaires for Standardizing Stroke Registry Data. Stud Health Technol Inform. 2024 Aug 22;316:1453–1457.
[7] Postcoordinated expression https://confluence.ihtsdotools.org/display/DOCGLOSS/postcoordinated+expression. [accessed 2025-01-06].
[8] Schulz S, Martínez-Costa C, Karlsson D, Cornet R, Brochhausen M, Rector A. An ontological analysis of reference in health record statements. Formal Ontology in Information Systems 2014 (267): 289–302.
[9] Rossander A, Lindsköld L, Ranerup A, Karlsson D. A state-of-the-art review of SNOMED CT terminology binding and recommendations. Methods Inf Med. 2021 Dec;60(S 02):e76–e88.

[2] This is why we have not used the FHIR resource "Questionnaire", which would only be appropriate in scenario limited to data exchange only.