Computer Methods and Programs in Biomedicine. 2017; 149: 1-9

SNOMED2HL7: a tool to normalize and bind SNOMED CT concepts to the HL7 Reference Information Model

D. Perez-Rey[a], R. Alonso-Calvo[a], S. Paraiso-Medina[a], C.R. Munteanu[b,c], M. Garcia-Remesal[a]

[a] *Biomedical Informatics Group, School of Computer Science, Universidad Politecnica de Madrid. Campus de Montegancedo, s/n, 28660, Boadilla del Monte, Madrid, Spain*
[b] *RNASA-IMEDIR, Computer Science Faculty, University of A Coruna, Campus de Elviña s/n, 15071, A Coruña, Spain*
[c] *Instituto de Investigación Biomédica de A Coruña (INIBIC), Complexo Hospitalario Universitario de A Coruña (CHUAC), A Coruña, Spain*

Abstract
Background. Current clinical research and practice requires interoperability among systems in a complex and highly dynamic domain. There has been a significant effort in recent years to develop integrative common data models and domain terminologies. Such efforts have not completely solved the challenges associated with clinical data that are distributed among different and heterogeneous institutions with different systems to encode the information. Currently, when providing homogeneous interfaces to exploit clinical data, certain transformations still involve manual and time-consuming processes that could be automated.
Objectives. There is a lack of tools to support data experts adopting clinical standards. This absence is especially significant when links between data model and vocabulary are required. The objective of this work is to present SNOMED2HL7, a novel tool to automatically link biomedical concepts from widely used terminologies, and the corresponding clinical context, to the HL7 Reference Information Model (RIM).
Methods. Based on the recommendations of the International Health Terminology Standards Development Organisation (IHTSDO), the SNOMED Normal Form has been implemented within SNOMED2HL7 to decompose and provide a method to reduce the number of options to store the same information. The binding of clinical terminologies to HL7 RIM components is the core of SNOMED2HL7, where terminology concepts have been annotated with the corresponding options within the interoperability standard. A web-based tool has been developed to automatically provide information from the normalization mechanisms and the terminology binding.
Results. SNOMED2HL7 binding coverage includes the majority of the concepts used to annotate legacy systems. It follows HL7 recommendations to solve binding overlaps and provides the binding of the normalized version of the concepts. The first version of the tool, available at http://kandel.dia.fi.upm.es:8078, has been validated in EU funded projects to integrate real world data for clinical research with an 88.47% of accuracy.
Conclusions. This paper presents the first initiative to automatically retrieve concept-centered information required to transform legacy data into widely adopted interoperability standards. Although additional functionality will extend capabilities to automate data transformations, SNOMED2HL7 already provides the functionality required for the clinical interoperability community.

Keywords
Interoperability; SNOMED CT; HL7; Normalization; Data integration

1. Introduction

Data sharing among biomedical researchers has produced important results in the past [1]. The inclusion of -omics-based biomarkers [2] and personalized medicine [3] have introduced new challenges in the area of biomedical research, and interoperability has been identified as one of the main issues for health care systems [4]. However, in such a complex and dynamic domain, interoperability standards are still not widely adopted. Semantic technologies were introduced to health care from a very different world [5], but with objectives similar to those of traditional interoperability standards. However, traditional semantic Web solutions usually rely on complete and sound ontologies, while there is a lack of such vocabularies and data models in biomedicine [6], [7].

Interoperability among health care systems has been extensively investigated over the last few years. Data models are the core components of such initiatives with heterogeneous solutions based on CDISC ODM [8], OMOP [9], i2b2 [10], OpenEHR [11] or Health Level 7 Reference Information Model (HL7 RIM) [12] among others. HL7 RIM version 3 provides an abstract model, based on clinical *Acts*, to represent clinical information. The objective is to build interoperability messages to communicate among information systems, including classes with sets of attributes associated by relationships. HL7 RIM is one of the most widely adopted standards, providing representation capabilities required to store complex, multi-scale and heterogeneous information.

Data models should be populated, through Extract Transform and Load (ETL) processes, with clinical concepts from biomedical terminologies such as SNOMED CT [13], NCI Thesaurus [14], ICD [15], LOINC [16] or HGNC [17] among others. The SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) is a logic-based clinical terminology, providing over 310,000 concepts and 1300,000 relationships among them. It is organized following a hierarchical structure of "is a" relationships—e.g., "Allergic asthma" is an "Asthma"—allowing for the composition of new concepts by a post-coordination mechanism. The extension of a controlled vocabulary generates ambiguity when representing information. To reduce such issues, the SNOMED CT includes a normalization mechanism that produces a Normal Form version of any given concept. Although they are not complete, such mechanisms are required to improve the sustainability of data management in current clinical research. In this framework, the data is frequently distributed among different and heterogeneous institutions, and it is impossible to force every institution to encode information in the same way (even using the same standard).

Once data has been annotated with a biomedical terminology, concepts should be linked to data model classes and attributes. Although binding from SNOMED CT to HL7-RIM has been previously documented [18], [19], [20], to the best of the authors' knowledge, there is a lack of automatic methods to provide such binding including HL7 recommendations and solving branch overlapping among terminologies and HL7 RIM. SNOMED2HL7 is a novel tool implementing such automatic method to normalize and bind SNOMED CT concepts to HL7 RIM classes. Additional information regarding the entire technical process of integrating clinical data with the proposed standards can be found elsewhere [21]. The objective of SNOMED2HL7 is to facilitate the binding between a comprehensive clinical terminology—i.e., SNOMED CT—and a widely adopted clinical interoperability standard—i.e., HL7 RIM—considering the context of the data source.

2. SNOMED normal form

Biomedical vocabularies containing hierarchical relationships frequently allow multiple options to code the same information. In SNOMED CT, such ambiguity is due to the existence of concepts that are the result of combining other (pre-coordinated) concepts—i.e., the SNOMED CT term "Nonvenomous insect bite of breast with infection" is a unique term with the code 15034009, but it involves three different concepts: (i) a nonvenomous insect bite, (ii) in the breast, (iii) that has produced an infection. Ideally, when integrating information from different heterogeneous sources, the data contained in the original repository should be decomposed and stored uniformly.

SNOMED CT provides a description of the process required to obtain the canonical concepts that compounds pre-coordinated concepts. The SNOMED Normal Form mechanism is a set of logical transformation rules that exploit the relationships present in SNOMED CT [22] and avoid possible redundancies in the vocabulary [23]. The input is a pre-coordinated concept, and the result is a set of concepts and relationships that represent the *Normal Form* of a concept. Concepts in the *Normal Form* are always labeled as *primitive*. A concept is *primitive* in SNOMED when its *roles*—i.e., attribute-value relationships with other concepts—and *parent* concepts do not fully express its meaning. Thus, *primitive* concepts provide a complete meaning by themselves, and any composition using other terms cannot fully represent the same meaning.

The most common *Normal Form* versions that have been defined for SNOMED CT are: (i) the *Long Normal Form* (LNF) and (ii) the *Short Normal Form* (SNF). The LNF contains every value of the concept *roles*; therefore, it contains every concept and attribute that can be inferred from the original term. The SNF only enumerates the *roles* that differentiate the concept from its primitive ancestors. According to [24], there is no loss of specificity when using the SNF compared to the LNF or the original term.

The Normal Form of any concept in SNOMED is an expression containing: (i) the *focus concept* and (ii) the *refinement*. The *focus concept* is the closest primitive from ancestors of the original concept. The *refinement* contains those *roles* that define the original concept, and it is composed of attribute-value pairs. Where the attribute is a relationship, and the value is a normalized expression (or a primitive concept). Fig. 1 depicts an example of the SNOMED Short Normal Form.

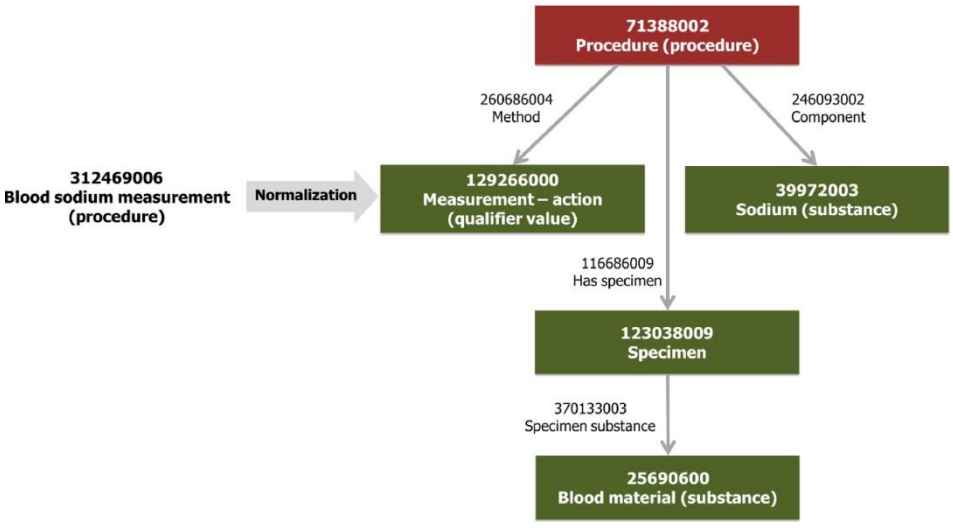

Fig. 1. Short Normal Form of "Blood sodium measurement".

The SNF from the concept “Blood sodium measurement (procedure)”, with SNOMED CT code 312469006, includes the focus concept “procedure” and three different roles: (i) a method (“Measurement action”), (ii) the corresponding specimen (the substance “blood material”) and (iii) the component (“Sodium”). The objective of this work is to provide an automatic tool that exploits the SNOMED Normal Form mechanism in conjunction with the terminology binding to the HL7 RIM.

3. Terminology binding to the HL7 Reference Information Model

When annotating new datasets with a controlled vocabulary, such as SNOMED CT, to be mapped to HL7 RIM components, clear links between data types and data model classes and attributes are required. The terminology binding defines which concepts can be stored in each field in the data model. Although SNOMED CT provides a wide coverage in clinical areas, and a large number of relationships, such wide coverage also hinders the definition of clear rules to automatically store data. Such links among vocabularies and data schema are particularly required with the existence of pre- and post-coordinated concepts. For instance, although “Malignant epithelial neoplasm of lung” is clearly defined by SNOMED CT as a *disorder*, if it is stored as an HL7 RIM *Act* containing the pre-coordinated version, this prevents queries from retrieving implicit information (i.e., the affected body part). This problem can be addressed by using the SNOMED CT normalization process together with terminology binding, where only primitive concepts are bound and stored in the data model—e.g., the concepts described in Fig. 1 should be stored as an *Act* representing a “Malignant epithelial neoplasm” with *Target Site* “Lung structure”.

The Terminfo project [25] has published a document describing how to use the relationships obtained in the Normal Form—i.e., *Finding Site, Method, Part of*—to select the link of a pre-coordinated SNOMED CT concept to an HL7 class. Nevertheless, there are other relationships that could appear in the Normal Form that are not included in the HL7 recommendations—i.e., Associated morphology, *Pathological Process*—and a binding to the RIM model is not defined in such cases. In addition, to create the terminology binding of SNOMED CT and HL7 RIM considering the ‘is a’ relationship (that defines a hierarchy), an assignment of the binding between SNOMED CT branches and attributes of HL7 RIM is required. Such hierarchical definition of the binding in SNOMED CT produces overlaps among branches—i.e., in SNOMED CT there are concepts with more than one parent that could belong to different branches. In these overlapping cases, the same concept could have different bindings to the HL7 model. Although most of the SNOMED concepts have a unique inherent context—from SNOMED branches mentioned above—other concepts can be used for more than one context (overlaps). SNOMED2HL7 can be used to select the appropriate context provided by each Electronic Health Record (EHR) following a normalized procedure.

In the present work, the terminology binding between SNOMED CT and HL7 has been implemented by annotating the ontology version of SNOMED CT. Annotations include the class and attribute from the HL7 RIM where each concept should be stored. The following sections are focused on the details of such implementation.

4. SNOMED2HL7 implementation

4.1. Architecture and technologies

Current approaches based on SNOMED CT and HL7 RIM to integrate clinical data rely on textual documentation for the ETL process. The SNOMED2HL7 tool aims to provide a means to automatically query a SNOMED CT concept normalization and terminology binding to the HL7 RIM. Fig. 2 describes the architecture and technologies used to provide a web application.

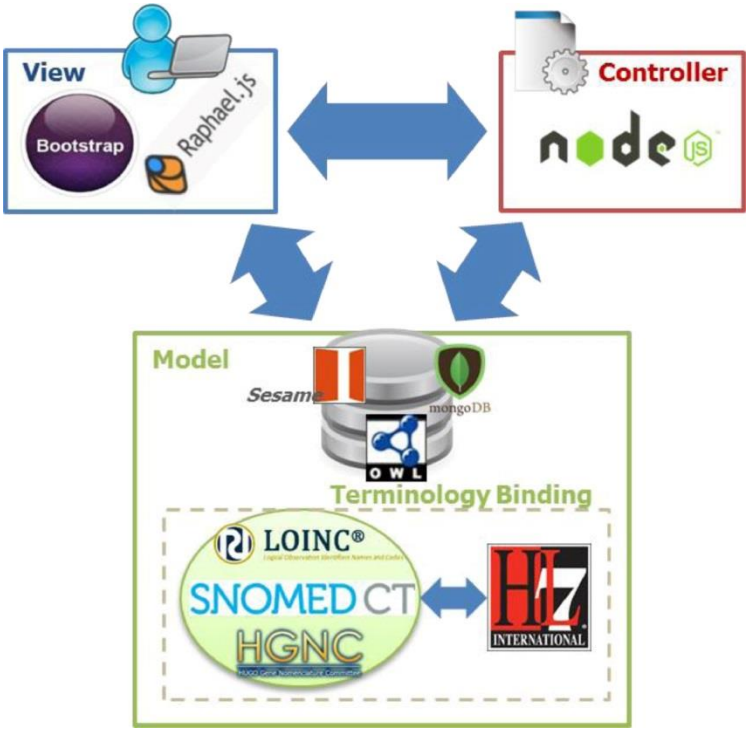

Fig. 2. SNOMED2HL7 implementation technologies.

The user interface was developed with Bootstrap [26] and Raphael [27] frameworks. Bootstrap, developed by twitter, is a web development framework comprised of a set of CSS and Javascript files for the web design. Therefore, it is used to define the complete web design and the format of the table. Raphael is a library for easily developing dynamic graphics and diagrams. In particular, the graphic representation of the SNOMED CT relationships and the HL7 RIM representation of the concept information have been implemented with the Raphael library. The controller side is governed by the Node.js platform [28]. It is an event architecture-based environment responsible for the scalability and interoperability of the different components on the server side.

The model side is responsible for managing, processing and storing application data. It includes technologies such as MongoDB [29] for access control and Sesame [30] for semantic resource management and representation of terminology binding. Such technologies, frameworks and semantic standards allowed for the development of a cross-platform accessible web application. They can also be deployed as SOAP and REST services to be used by third parties. Terminology binding vocabularies and related technologies are described in the following sections.

4.2. Terminology binding representation

The representation of large terminologies such as SNOMED CT using semantic technologies is a critical step in the process of developing SNOMED2HL7 due to performance requirements. An OWL file with the SNOMED CT release has been generated using the official script provided with the tab-delimited release annotated with additional binding information. Additionally, mapping information to HL7 RIM is included for every SNOMED CT concept following the Provenance Vocabulary Core Ontology specification [31].

The terminology binding process analyzes the position of the concept within the SNOMED CT hierarchy to determine the mapping to HL7 RIM. Below is the code generated for the "Blood sodium measurement (procedure)" concept with its relations, synonyms among other attributes.

```
<owl:Class rdf:about="SCT_312469006">
<rdfs:label xml:lang="en">Blood sodium measurement (procedure)</rdfs:label>
<skos:altLabel xml:lang="en">Sodium measurement, blood</skos:altLabel>
<skos:altLabel xml:lang="en">Blood sodium level</skos:altLabel>
<skos:altLabel xml:lang="en">Blood sodium measurement</skos:altLabel>
≤prv:containedBy rdf:resource="http://www.gib.fi.upm.es/
hl7rim-common-data-model/# Procedure_code"/≥
≤prv:containedBy rdf:resource="http://www.gib.fi.upm.es/
hl7rim-common-data-model/# Procedure_methodCode"/≥
≤prv:containedBy rdf:resource="http://www.gib.fi.upm.es/
hl7rim-common-data-model/# Observation_code"/≥
≤prv:containedBy rdf:resource="http://www.gib.fi.upm.es/
hl7rim-common-data-model/# Observation_methodCode"/≥
<rdfs:subClassOf rdf:resource="SCT_25197003"/>
<rdfs:subClassOf rdf:resource="SCT_166932001"/>
<owl:intersectionOf rdf:parseType="Collection">
<owl:Class rdf:about="SCT_25197003"/>
<owl:Class rdf:about="SCT_166932001"/>
<owl:Restriction>
<owl:onProperty rdf:resource="RoleGroup"/>
<owl:someValuesFrom>
<owl:Restriction>
<owl:onProperty rdf:resource="SCT_246093002"/>
<owl:someValuesFrom rdf:resource="SCT_39972003"/>
</owl:Restriction>
</owl:someValuesFrom>
</owl:Restriction>
<owl:Restriction>
<owl:onProperty rdf:resource="RoleGroup"/>
<owl:someValuesFrom>
<owl:Restriction>
```

```
<owl:onProperty rdf:resource="SCT_116686009"/>
<owl:someValuesFrom rdf:resource="SCT_119297000"/>
</owl:Restriction>
</owl:someValuesFrom>
</owl:Restriction>
<owl:Restriction>
<owl:onProperty rdf:resource="RoleGroup"/>
<owl:someValuesFrom>
<owl:Restriction>
<owl:onProperty rdf:resource="SCT_260686004"/>
<owl:someValuesFrom rdf:resource="SCT_129266000"/>
</owl:Restriction>
</owl:someValuesFrom>
</owl:Restriction>
</owl:intersectionOf>
</owl:Class>
```

The following OWL labels were used:

- ***owl:Class***: represents the class information for the concept "SCT_312469006"
- ***rdfs:label***: is the label of the concept, "Blood sodium measurement (procedure)"
- ***skos:altLabel***: is the set of alternative labels for the concept
- ***prv:containedBy***: is the terminology mapping of the concept in the HL7 RIM
- ***rdfs:subClassOf***: hierarchical information of the concept
- ***owl:intersectionOf***: statement for describing individuals who are members of the class description. This example contains a set of *owl:Restriction* to represent a property restriction. Every restriction contains an *owl:onProperty* to represent the relationships of the concepts—e.g., "SCT_116676009" is the "Has Specimen" relationship and the *owl:someValuesFrom* statement to represent the values associated with the relationship and "SCT_119297000" is the "Specimen".

SNOMED CT provides a large coverage to represent the entire clinical domain; however, certain domains are traditionally covered by overlapping terminologies [32]. Since SNOMED CT has a limited coverage for certain data types, in the current version of SNOMED2HL7, LOINC and HGNC terminologies have been also included. LOINC has been bind to laboratory tests (HL7 RIM Observation) due to the adoption of LOINC as "de facto" standard in the area. HGNC has been bind to RIM entities and it is accessible through the "Gene" radio button to increase the coverage for gene names. As they cover different data types, mapping among SNOMED CT, LOINC and HGNC was not required. The resulting OWL file contains more than 7 million tags stored in a Sesame repository.

4.3. User interface

The first step to normalize SNOMED CT concepts and bind to HL7 RIM is to find the requested concept. The initial interface of SNOMED2HL7 is a text search with autocomplete to suggest concept strings. Fig. 3 shows the initial interface of SNOMED2HL7.

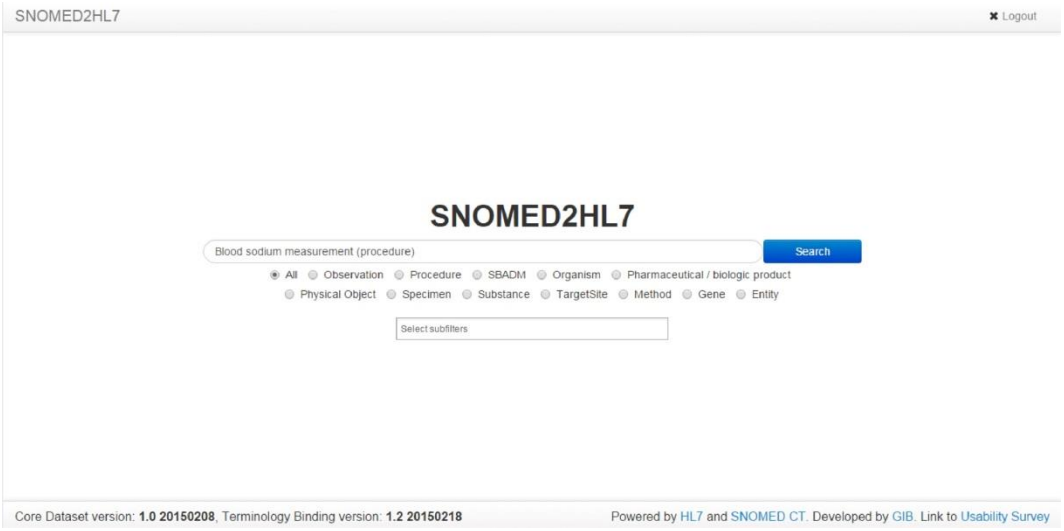

Fig. 3. SNOMED2HL7 webpage.

A set of filters are defined to restrict searches within a given context. These filters are based on the general classes of the HL7 RIM—i.e., *Observation*, *Procedure*, *Substance Administration*, *Entity*, *Method* and *Target Site*. When selected, the autocomplete service only provides concepts within this context. Due to the broad scope of the *Entity* filter, it is divided into a set of more specific filters: *Organism*, *Pharmaceutical* or *Biologic product*, *Physical object*, S*pecimen*, S*ubstance* and G*ene*. Once a concept is selected, SNOMED2HL7 provides the information presented in Fig. 4.

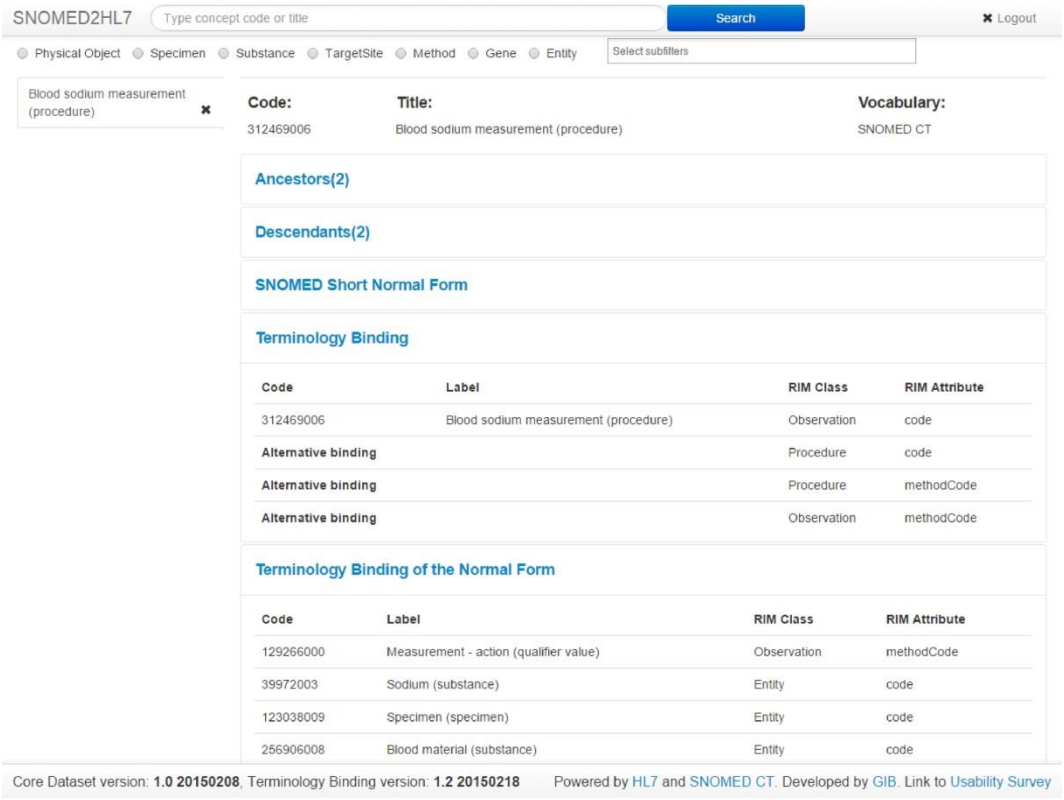

Fig. 4. Blood sodium measurement (procedure) in SNOMED2HL7.

The following information is provided by SNOMED2HL7 for each SNOMED concept:

- **General information:** concept code and full name
- **Ancestors and descendants**: concept codes and names of parents and children of the concept in the SNOMED CT hierarchy
- **Normal Form**: short normal form of the selected concept, represented by a table with the focus concept, relationships and relationship values
- **Terminology Binding**: mapping representation of the original concept to HL7 RIM classes and attributes
- **Terminology Binding of the normal form**: mapping representation of the short normal form of the original concept to HL7 RIM classes and attributes
- **HL7 RIM Representation**: graphical representation of the concept in the HL7 RIM model.
- In case of LOINC concepts, SNOMED2HL7 provides the LOINC properties of the concept (time aspect, system, scale type, method type, class, short and long name). With this information, SNOMED2HL7 allows for browsing of the SNOMED CT vocabulary to retrieve concept-related information and corresponding HL7 RIM mappings.

5. Results

SNOMED2HL7 has been developed as a web application, and it is available at http://kandel.dia.fi.upm.es:8078. It contains bindings to RIM classes and attributes with the corresponding SNOMED Normal Form for 94.5% of SNOMED CT concepts (SnomedCT_Release_INT_20160731 version). To support specific data annotations, additional terminologies have been included that bind to fixed HL7 classes: (i) for laboratory tests, the 2.45 release of LOINC has been bound to *Observation*; and (ii) for gene names, the 2013 release of HGNC has been bound to the *Entity* class. The following sections describe the SNOMED CT binding coverage of SNOMED2HL7 and the binding overlaps among SNOMED CT.

5.1. SNOMED CT binding coverage

SNOMED2HL7 follows HL7 recommendations to link SNOMED CT concepts to HL7 RIM classes [25]. The following table shows the total number of concepts linked to each of the main classes: (i) observation, (ii) procedure, (iii) entity and (iv) substance administration. The table also shows the total number of concepts linked to each of the main modifiers: (i) method, (ii) target site and (iii) approach site. Table 1 contains statistics on classes assigned to SNOMED CT concepts and modifiers by SNOMED2HL7.

Table 1. Distribution of SNOMED CT concepts linked to each HL7 RIM class.

HL7 RIM class	# concepts	% distribution
Observation	138,624	37.48
Procedure	55,960	15.12
Entity	92,148	24.90
Substance Administration	2058	0.56
Method	54,512	14.74
Target Site	26,303	7.11
Approach Site	329	0.09

With a total number of 450 K bindings and almost 300 K unique concepts mapped (target site and method bindings are duplicated for both observation and procedure), *Observation* is the most common HL7 RIM class for SNOMED CT concepts, with more than one-third of the concepts, followed by *Entities* and *Procedures* and the modifiers *Method* and *Target Site*. Table 2 contains the number of concepts and coverage per SNOMED CT branch mapped in SNOMED2HL7.

Table 2. Terminology binding coverage.

SNOMED CT Branch	# concepts	% coverage
Body Structure	30,244	99.86%
Clinical Finding	97,167	100.00%
Environment or Geographical Location	1696	100.00%
Event	3652	100.00%
Observable Entity	8215	100.00%
Organism	31,926	100.00%
Pharmaceutical Biologic Product	16,761	100.00%
Physical Object	4510	100.00%
Procedure	52,692	100.00%
Qualifier Value	8889	8.64%
Situation with Explicit Context	3194	97.93%
Specimen	1329	100.00%
Substance	23,674	100.00%

Most of the branches are fully covered by SNOMED2HL7. The *Body Structure* and *Situation with Explicit Context* branches are mostly covered but include: (i) 43 concepts from *Body Structure* referring to cell interactions and generic concepts—e.g., 21229009 "Topography not assigned"—and (ii) 66 concepts from *Situation with Explicit Context* referring to exceptional situations in the patient context—e.g., 160861007 "Spouse arrested (situation)"—are not mapped.

For the *Qualifier Value* branch, only 8.64% of concepts (768 out of 8889) have a mapping, including 328 concepts as *Approach Site*, 434 as *Method* and 6 as *Target Site*. Most *Qualifier Value* concepts lack a complete meaning without being combined with other concepts. They must be associated with an attribute concept to create an attribute-value pair—e.g., in the attribute-value pair "laterality-left", the *Qualifier Value* "left" has no meaning without the attribute "laterality".

Finally, because they are not frequently used in ETL processes, other SNOMED CT branches—i.e., Physical Force, Record Artifact, SNOMED CT *Model Component*, Social Context, Special Concept, and *Staging and Scales*—are not mapped to HL7 RIM within the current SNOMED2HL7 version.

5.2. Binding overlaps among SNOMED CT branches

Automatic methods to bind SNOMED CT with HL7 RIM classes need to deal with multiple options to store the same concepts. SNOMED2HL7 contains 19,781 SNOMED CT concepts that could be bound to more than one HL7 RIM class and attribute. SNOMED CT binding overlaps are introduced by the hierarchical structure of SNOMED CT, where ancestors of a concept are bound to a given HL7 RIM class, while a portion of its offspring are bound to a different class.

As can be observed in Fig. 5, the majority of such overlaps, 17,837 (90.17%), are concepts that can be simultaneously a *Procedur*e or an *Observation*—e.g., [103800002] "Coagulation factor assay". Most of the remaining overlaps, 2058 (10.40%), are concepts that could be *Procedures* or *Substance Administration*—e.g., [9935002] "Local anesthesia surface by refrigerant". A total of 114 (0.57%) of the latest also include a small number of cases with a triple *Procedure* / *Observation* / *Substance Administration* overlapping—e.g., [252922007] "Injection of sentinel lymph node using ultrasound guidance". Additionally, because all root concepts of *Method* are

located in sub-branches of *Procedure*—e.g., 260686004| "Method" or 418775008| "Finding method" —all concepts bound to *Method* are also bound to *Procedure*, while 1896 concepts bound to *Procedure* are not bound to *Method*.

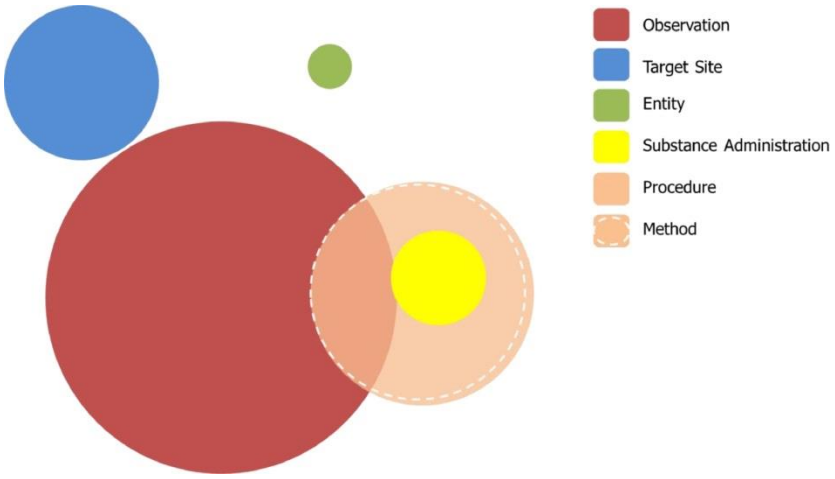

Fig. 5. Graphical distribution of binding overlaps covered by SNOMED2HL7.

To facilitate the selection of the appropriate binding depending on the context, SNOMED2HL7 provides alternatives when a binding overlap occurs—e.g., "Blood sodium measurement" can be bound to Observation, *Observation Method*, Procedure and *Procedure Method*. Because this concept may have values/units, SNOMED2HL7 suggests *Observation* as the preferred binding and includes the corresponding binding of the normalized form.

The upper section of Fig. 6 lists the four alternatives provided in the "Terminology Binding" section of SNOMED2HL7 to bind "Blood Sodium Measurement" to HL7 RIM components. The lower section of the figure shows the normalization of the preferred binding that can be produced with information from the "Terminology Binding of the Normal Form" section. Such normalization is used to enrich data in the process of mapping legacy repositories into HL7 RIM and SNOMED CT.

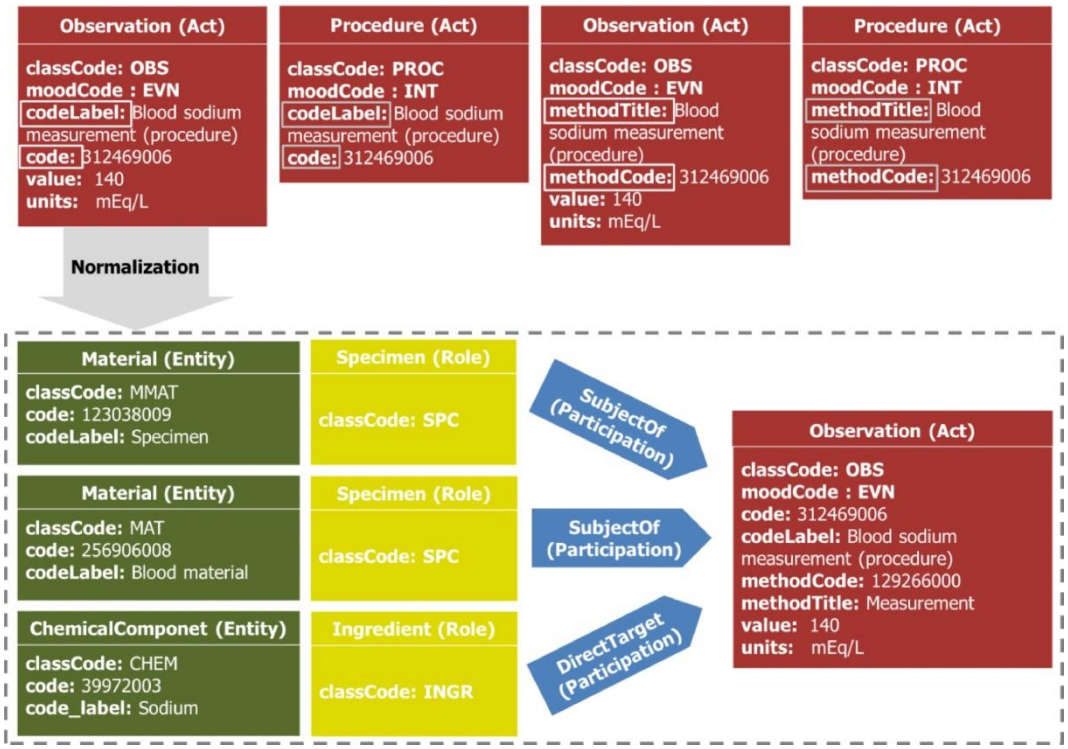

Fig. 6. SNOMED2HL7 binding options to HL7 for "Blood Sodium Measurement".

5.3. SNOMED2HL7 binding evaluation

To evaluate how accurately SNOMED2HL7 binds terminology concepts to HL7 RIM classes and attributes, we analyzed 24 datasets with more than 20.000 patients integrated within the EURECA project. These datasets have been stored in HL7 RIM data model, containing over 1.200 non repeated terminology concepts as presented in Table 3.

Table 3. SNOMED2HL7 evaluation in the EURECA project.

	HL7 RIM	*SNOMED2HL7 Preferred Binding*	*%*	*SNOMED2HL7 Alternative Binding*	*%*	*Total*
Entity	407	403	99,02%	0	000%	99,02%
Observation	519	475	91,52%	0	0,00%	91,52%
Procedure	65	62	95,38%	3	4,62%	100,00%
Substance Administration	8	6	75,00%	0	0,00%	75,00%
Target Site	121	104	85,95%	0	000%	85,95%
Interpretation	5	0	000%	0	000%	000%
Method	25	0	000%	14	56,00%	56,00%
Observation Values	56	0	000%	0	000%	000%
Total	**1206**	**1050**	**87,06%**	**17**	**1,41%**	**88,47%**

This table shows how terminology concepts are stored, compared to the binding proposal of SNOMED2HL7, while the corresponding set of concepts is available through the application website. Since datasets were focused in oncology patients, the majority of the concepts are represented with *Observation*, *Procedure* and *Entity* RIM classes as can be seen in "*HL7 RIM*" column. The "*Preferred Binding*" column represents the amount of such concepts that are correctly bind with the first binding option suggested by SNOMED2HL7 (preferred). Among the rest of the concepts, the "*Alternative Binding*" column represents the number of concepts that are correctly bind with any of the alternative bindings produced by branch overlaps.

6. Discussion and conclusions

Since most of the concepts provide only a preferred bind (without alternatives), the accuracy is only improved from 87,06% to 88,47% with alternative bindings. They are however needed for specific sections such as Method, with a high impact of alternative bindings. The most relevant classes (Entity, Observation and Procedure) produced such high accuracy, while the majority of the 11,53% of concepts without correct SNOMED2HL7 bindings were in fact annotated with other clinical terminologies, such as ICD10, MeSH and NCI Thesaurus, that will be included in SNOMED2HL7 in the future.

Currently, SNOMED CT and HL7 RIM are two of the most powerful and widely adopted standards in clinical interoperability. Normalization mechanisms such as the SNOMED Normal Form reduce the ambiguity of a complex biomedical domain, where even within a controlled vocabulary there are different options to represent the same information. Until now, the process of deciding which HL7 RIM class is related to each SNOMED CT has been manually performed by browsing extensive documentation. To the knowledge of the authors and experts evaluating SNOMED2HL7, there were no automatic tools to provide such information.

SNOMED2HL7 provides information on the process of data annotation for selecting the appropriate concept. Although annotation cannot be fully automated, SNOMED2HL7 facilitates the task of selecting the corresponding HL7 RIM component including the clinical context of the concept in the process. Normalization mechanisms have been implemented in SNOMED2HL7 to reduce ambiguity when mapping legacy repositories into HL7 RIM components or generating HL7 messages.

Centered on SNOMED CT, additional terminologies have been included in SNOMED2HL7 for specific data types. LOINC for laboratory tests and HGNC for gene names have been included in the current version. To increase the coverage of legacy systems, further versions of the tool will include new terminologies. Automatic generation of HL7 message templates and dedicated APIs to automatically bind a set of concepts will also be provided in future versions to facilitate the ETL process.

The results show that SNOMED2HL7 provides a high coverage of the domain, including mechanisms to solve binding overlaps, and around 90% of binding accuracy after an evaluation with real datasets. Although additional functionality is still required to advance in the process of integrating clinical legacy systems, initial evaluations suggest that the current status of SNOMED2HL7 provides the functionality required for the clinical interoperability community.

Summary points

What is already known about this topic?

- Clinical common data models and terminologies has been extensively developed in recent years
- However, the process of transforming clinical data to ensure interoperability still requires significant manual effort
- There is a lack of automatic methods to link data models and terminologies
- What this study added to our knowledge?
- This work proved that such automatic methods can be developed for major data models and terminologies such as SNOMED CT and HL7 RIM
- Normalization methods centered on concepts to reduce ambiguity can be also integrated in such automatic solutions
- With a coverage close to 100% for the most relevant sections

Authors' contributions

Perez-Rey and Alonso-Calvo conceived the study, conceptualized the design and were responsible for the organization and creation of the manuscript. ParaisoMedina led the implementation of the tool and contributed to the manuscript. Garcia-Remesal and Munteanu contributed to the evaluation and provided a critical revision of the manuscript. All authors read and approved the final manuscript.

Declaration of conflicting interests

The authors have no conflicts of interest to declare.

Acknowledgments

This work was supported by the CIMED collaborative project cofunded by ISCIII and FEDER under the grant number PI13/02020 and by the European Commission through EURECA (FP7-ICT-2011-7-288048) project.

References

[1] O. Abe, R. Abe, K.T. Enomoto, K. Kikuchi, H. Koyama, Y. Nomura, K. Sakai. Tamoxifen for early breast cancer: an overview of the randomised trials. Lancet, 351 (9114) (1998), pp. 1451-1467

[2] L.M. McShane, M.M. Cavenagh, T. Lively, D.A. Eberhard, W.L. Bigbee, M.P. Williams, J.P. Mesirov, M.Y. Polley, K.Y. Kim, J.V. Tricoli, *et al.* Criteria for the use of omics-based predictors in clinical trials. Nature, 502 (2013), pp. 317-320

[3] M.A. Hamburg, F.S. Collins. The path to personalized medicine. New Engl. J. Med., 363 (4) (2010), pp. 301-304

[4] D.J. Brailer. Interoperability: the key to the future health care system. Health Affairs-Millwood Va Then Bethesda Ma, 24 (2005), p. W5

[5] C. Kei-Hoi, E. Prud'hommeaux, Y. Wang, S. Stephens. Semantic web for health care and life sciences: a review of the state of the art. Briefings Bioinform., 10 (2) (2009), pp. 111-113

[6] J. Ingenerf, R. Linder. Assessing applicability of ontological principles to different types of biomedical vocabularies. Methods Inf. Med., 48 (5) (2009), p. 459

[7] C.N. Mead. Data interchange standards in healthcare it-computable semantic interoperability: now possible but still difficult. do we really need a better mousetrap?. J. Healthcare Inf. Manage., 20 (1) (2006), p. 71
[8] W. Kuchinke, J. Aerts, S.C. Semler, C. Ohmann. CDISC standard-based electronic archiving of clinical trials. Methods Inform. Med., 48 (5) (2009), p. 408
[9] P.E. Stang, P.B. Ryan, J.A. Racoosin, J.M. Overhage, A.G. Hartzema, C. Reich, E. Welebob, T. Scamecchia, J. Woodcock. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. In Ann. Intern. Med., 153 (9) (2010 Nov), pp. 600-606
[10] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, C.C. Henry, S. Churchill, I. Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J. Am. Med. Inf. Assoc., 17 (2) (2010), pp. 124-130
[11] S. Garde, K. Petra, E.J.S. Hovenga, S. Heard. Towards semantic interoperability for electronic health records–domain knowledge governance for open EHR archetypes. Methods Inf. Med., 46 (3) (2007), pp. 332-343
[12] G. Beeler, J. Case, J. Curry, A. Hueber, G. Mckenzie LSchadow, A.M. Shakir. "HL7 reference information model". 2003.
[13] K Donnelly. SNOMED-CT: the advanced terminology and coding system for eHealth. Med. Care Compunetics 3 (2006), pp. 279-290
[14] N. Sioutos, C. Sherri de, W.H. Margaret, W.H. Frank, S. Wen-Ling, W.Wright Lawrence. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J. Biomed. Inf., 40 (1) (2007), pp. 30-43
[15] International Classification of Diseases (ICD)- World Health Organization. http://www.who.int/classifications/icd/en/ (2017)
[16] C.J. McDonald, S.M. Huff, J.G. Suico, G. Hill, D. Leavelle, R. Aller, A. Forrey, K. Mercer, G. DeMoor, J. Hook, W. Williams, J. Case, P. Maloney. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clinical Chem., 49 (4) (2003), pp. 624-633
[17] R.L. Seal, S.M. Gordon, M.J. Lush, M.W. Wright, E.A. BrufordGenenames.org: the HGNC resources in 2011. Nucleic Acids Res., 39 (2011), pp. 514-519
[18] S. Heymans, M. McKennirey, J. Phillips. Semantic validation of the use of SNOMED CT in HL7 clinical documents. J. Biomed. Semant., 2 (2011), p. 2
[19] Rico-Diez, S. Aso, D. Perez-Rey, R. Alonso-Calvo, A. Bucur, B. Claerhout, V. Maojo. SNOMED CT normal form and HL7 RIM binding to normalize clinical data from cancer trials. Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on, IEEE (2013), pp. 1-4
[20] S. Paraiso-Medina, D. Perez-Rey, A. Bucur, B. Claerhout, R. Alonso-Calvo. Semantic normalization and query abstraction based on SNOMED-CT and HL7: supporting multicentric clinical Trials. IEEE J. Biomed. Health Inf., 19 (3) (2015), pp. 1061-1067
[21] EURECA ETL Guidelines. Seen at https://github.com/gib-upm/Clinical-Trial-Semantic-Interoperability-Solution/blob/master/docs/ETL_Guidelines.pdf (2017)
[22] K.A. Spackman. Normal forms for description logic expressions of clinical concepts in SNOMED RT. Proceedings of the AMIA Symposium, American Medical Informatics Association (2001), p. 627
[23] K. Dentler, R. Cornet. Intra-axiom redundancies in SNOMED CT. Artif. Intell. Med., 65 (1) (2015), pp. 29-34
[24] K.A. Spackman. Normal forms for description logic expressions of clinical concepts in SNOMED RT. Proc AMIA Symp. (2001), pp. 627-631
[25] HL7 TermInfo Project. http://www.hl7.org/special/committees/terminfo/ (2017)
[26] Twitter, Inc. Bootstrap framework homepage. http://twitter.github.io/bootstrap/ (2017)
[27] Raphael.js Homepage. Seen at http://raphaeljs.com/ (2017).
[28] Node.js Homepage. Seen at http://nodejs.org/ (2017)
[29] K. Chodorow, MongoDB: the definitive guide. O'Reilly Media, Inc. 2013.
[30] J. Broekstra, A. Kampman, F. Van Harmelen. Sesame: a generic architecture for storing and querying rdf and rdf schema. The Semantic Web—ISWC 2002, Springer, Berlin Heidelberg (2002), pp. 54-68
[31] O. Hartig, J. Zhao, Provenance vocabulary core ontology specification. Latest version available at 2012. http://trdf.sourceforge.net/provenance/ns.html
[32] O. Bodenreider. Issues in mapping LOINC laboratory tests to SNOMED CT. AMIA Annual Symposium Proceedings, 2008, American Medical Informatics Association (2008), p. 51