Formal representation of disorder associations in SNOMED CT

Edward Cheetham[1], Yongsheng Gao[2], Bruce Goldberg[3], Robert Hausam[4], and Stefan Schulz[5,*]

[1] Health and Social Care Information Centre, UK
[2]International Health Terminology Standards Development Organisation, Copenhagen, Denmark
[3]Kaiser Permanente, USA
[4]Hausam Consulting LLC, Midvale, UT, USA
[5] Institute of Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria

ABSTRACT

Medical terminologies like SNOMED CT often provide codes for frequently co-occurring associations of findings and disorders, such as syndromes or diseases with sequelae. The current release of SNOMED CT still lacks a principled solution for representing these concepts, which was the reason for the IHTSDO project group "Event, Condition, Episode" to elaborate a well-founded approach based on criteria of formal ontology. The group analysed complex SNOMED CT terms and proposes a simple solution, which draws on the interpretation of findings, disorders, and diseases as clinical life phases. Co-occurrence, temporal relatedness and causal relatedness were represented by distinct modelling patterns in OWL-DL.

1 INTRODUCTION

A main purpose of clinical terminologies is to support semantic annotation of the content of medical records. Consequently, in many terminology systems such as ICD-9 and ICD-10, in the draft of the upcoming ICD-11 (WHO, 2015), as well as in SNOMED CT (IHTSDO, 2015), numerous codes denote clinical phenomena that frequently co-occur or are temporally related, so that complex disorders like *Pericarditis with pericardial effusion*, or *Vitamin B12 deficiency anaemia due to malabsorption* can be encoded in one step. Extreme cases are codes for highly specific clinical scenarios like *Extradural haemorrhage following injury without open intracranial wound and with prolonged loss of consciousness (more than 24 hours) without return to pre-existing conscious level*.

Table 1. English UMLS terms containing coordinating and temporal connectors, related to all English UMLS terms.

Substring	Count	Rate
' after '	3,899	0.1%
' and '	337,706	4.7%
' caused by '	3,605	0.0%
' due to '	29,223	0.4%
' with '	231,128	3.2%
' without '	21,131	0.3%
ALL	626,692	8.7%

[*] To whom correspondence should be addressed: stefan.schulz@medunigraz.at

A review of all English terms in the UMLS Metathesaurus (NLM, 2015), which constitutes the biggest collection of biomedical terminology, reveals the importance of coordinating expressions or particles as parts of domain-specific terms (Table 1). Most of these terms denote findings, events, disorders, and procedures. In SNOMED CT, the picture is similar, as shown in Table 2. Many of these terms had been incorporated into SNOMED CT because of efforts to align ICD 9 and 10 with SNOMED CT.

Table 2. Distribution of coordinating and temporal connectors in English SNOMED CT fully specified names (percentages, rightmost column absolute count). Total number of concepts approx. 300,000.

	' after '	' and '	' caused by '	' ue to'	'with'	'without'	Σ	Σ_{abs}
Body Struct.	.0	2.5	.0	.0	.6	.0	3.1	951
Clin. Finding	.1	3.0	.1	2.7	4.7	1.3	11.9	11,974
Event	.6	8.1	12.0	1.9	9.6	3.1	44.3	1,627
Obs. Entity	.3	2.3	.0	.0	.5	.0	3.1	257
Product	.0	1.5	.0	.0	.0	.0	1.5	259
Phys. Object	.0	2.3	.0	.0	5.4	1.2	9.0	408
Procedure	.1	5.8	.0	.0	5.8	.3	12.1	6,497
Qual. Value	.2	1.1	.0	.0	.4	.0	1.8	162
Situation	.3	1.8	.0	.2	3.9	.4	6.6	243
Substance	.0	1.1	.0	.0	.4	.0	1.5	353
Others	.0	1.2	.0	.0	.1	.0	1.3	584
ALL	.0	2.9	.2	1.0	3.0	.5	7.7	23,039

While SNOMED CT is increasingly incorporating principles of applied ontology and provides a description logics (DL) (Baader *et al.*, 2007) based version implementing OWL EL (Motik *et al.*, 2012), the current representation of co-ordinating expressions in SNOMED CT does not follow clearly defined patterns. For instance, the definition of the SNOMED CT concept *Diabetic retinopathy* (disorder) uses the relation **associated with** for linking with *Diabetes mellitus*, whereas *Paraneoplastic neuropathy* (disorder) is connected to *Neoplastic disease* using **due to**. Another example is given by the concepts *Dermatomycosis associated with AIDS* (disorder) and *AIDS with dermatomycosis* (disorder), which appear to be duplicates. Whereas the former one uses the relation **associated with** for establishing a connection with the concept *AIDS*, the latter one is represented as a subclass of *AIDS*. This motivated the project group *Event, Condition, Episode Model* (ECE) of IHTSDO[1], the organization that maintains SNOMED CT, to conduct a thorough investigation of this phenomenon and to suggest a solution that is in line with current principles of ontology development in SNOMED CT.

2 METHODS

The ECE group decided to limit the scope of the investigation to the SNOMED CT hierarchy *Clinical Finding / Disorder*, following IHTSDO's current strategic directions in the content development process (IHTSDO, 2010). SNOMED CT statements that implicitly include negation were also not considered because they are not expressible in OWL-EL. All group members selected SNOMED CT term samples that represented coordination phenomena, in order to propose recurring modelling patterns. Having done this, the group discussed the underlying meaning, in particular the ontological commitment of the sample *Finding / Disorder* concepts and the underlying semantics with regard to time and causality. As an ontological reference, BioTopLite2 (BTL2) (Schulz & Boeker, 2013), an upper-level ontology based on OWL DL and tailored for the biomedical domain, was used. BTL2 provides a small set of upper-level classes, mappable to BFO (2015). All BTL2 classes exhibit a set of constraining axioms using a set of canonical relations, partly derived from the OBO Relation Ontology (Smith *et al.*, 2005). BTL2 heavily constrains the freedom of the ontology engineer, which warrants a higher predictability of the ontologies produced.

Table 3. Four patterns found for "*X* with *Y*"

Pattern	Definition	Example
1	Both *X* and *Y* are co-occurrent, but with no causality or manifestational relationship between *X* and *Y*	*Hay fever with asthma*
2	*X* is due to *Y*, but *X* and *Y* are not necessarily co-occurrent	*Disorder of optic chiasm due to non-pituitary neoplasm*
3	X temporally follows Y. This does not specify that *X* is due to *Y*, although causality is frequently implied	*Postvaricella encephalitis*
4	X is due to *Y*, and both *X* and *Y* are co-occurrent	*Hernia, with intestinal obstruction*

[1] http://www.ihtsdo.org/participate/project-groups

3 RESULTS

3.1 Typology and ontological analysis

Our analysis yielded four patterns of coordinative expressions in the SNOMED CT Finding / Disorder hierarchy, as shown in Table 3.

Further analysis focused on the following questions:

- Which are exactly the entities that are denoted by the concepts under scrutiny?
- Which temporal relationships have to be distinguished?
- What does causality mean and how is it linked to temporality?

According to Schulz *et al.* (2012b), we interpret all finding / disorder codes as *Clinical Situations* or *Clinical Life Phases* (we will use the latter term and illustrate it by the suffix "CLP") i.e. a patient's life phase during which a clinically relevant condition is present. For instance, the SNOMED CT concept $\textit{Encephalitis}_{\text{CLP}}$ denotes the class of processual entities of the type *Life phase*, in which some encephalitis process is present in any temporal instant covered by this life phase. Accordingly, $\textit{Hernia}_{\text{CLP}}$ denotes the class of processual entity of the type *Life phase*, in which the material disorder *Hernia* is fully present. The advantage of this interpretation is that we do not have to deal with hierarchies of entity types of different ontological categories under the same umbrella. To this end, BTL2 provides the defined class *Condition* – the disjunction of *Material object, Disposition* and *Process* (Schulz *et al.*, 2011a) – and the class *Situation* as a life phase during which some condition holds: an X_{CLP} is a Life phase during which some condition *X* is fully present. If John has constant headache today from 6am to 11pm, this period of his life is of the type $\textit{Headache}_{\text{CLP}}$. If he is seen by a doctor between 3pm and 3.10pm, this ten-minute lifespan is a new instance of the same type. If he also suffers from diabetes mellitus, then these life phases also instantiate '*Diabetes mellitus*'$_{\text{CLP}}$. We formalize this in OWL DL in the following way, using OWL Manchester Syntax:

X_{CLP} equivalentTo '*Clinical life phase*' and
'**has condition**' some *X*

The relation '**has condition**' in BTL2 holds between a life phase and an entity that is constantly present during this life phase. Independently, we have to look at temporality, where we need to clarify what "following" and "co-occurring" exactly mean. In BFO2 we find the relation '**is preceded by**', which is defined as relating two processes, one of which ends before the second one begins.

A commonly accepted framework for describing temporal relations is Allen's (1983) interval calculus (Fig. 1). Compared to this, a relational statement based on BTL2 "x **is preceded by** y" corresponds to either the Allen-based statement "y **takes place before** x" or "y **meets** x".

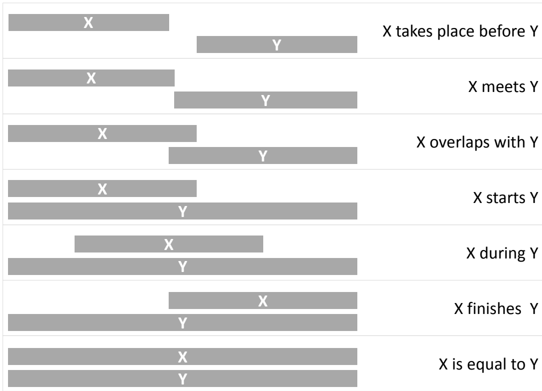

Fig. 1. Base relations of Allen's interval calculus. The converse relations are not depicted

Looking at the examples where we had asserted co-occurrence, we agreed to interpret "x **co-occurs with** y" as the disjunction of "x **starts** y", "x **during** y", "x **finished** y", and "x **is equal to** y". Let us take the example *Hay fever with asthma* (Fig. 2). We have three *Clinical life phase* entities: '*Hay fever with asthma*'$_{\text{CLP}}$, '*Hay fever*'$_{\text{CLP}}$, and *Asthma*$_{\text{CLP}}$. The possible temporal patterns result from any combination of Fig.2 left hand side with Fig.2 right hand side. All temporal instants of Hay fever with *Asthma*$_{\text{CLP}}$ temporally coincide with some instant of '*Hay fever*'$_{\text{CLP}}$ and some instant of *Asthma*$_{\text{CLP}}$.

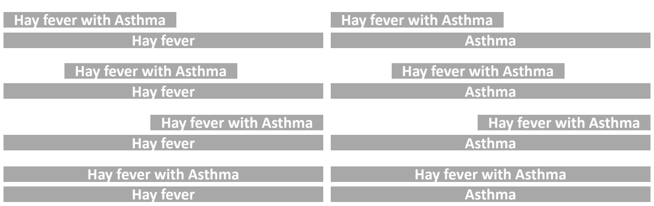

Fig. 2. Co-occurrence of combined situations

Finally, we will have a look at causality. Notwithstanding past and on-going philosophical debates about its nature, we consider the notion of causality as a primitive predicate, which is essential for medical reasoning and decision-making. Whether y follows x accidentally or because it is caused by x is seen as fundamentally different. There are important temporal implications of causality. It is a truism that an effect cannot precede its cause, or conversely, an effect has to follow its cause. Referring to the Allen calculus, "x **causes** y" would then be only compatible with "x **takes place before** y", "x **meets** y", "x **overlaps** y" as well as with (switching the arguments x and y) "y **during** x" and "y **finishes** x". All these relations have in common that the starting point of x precedes the starting point of y.

3.2 Proposal of modelling patterns

In the following, we will propose modelling solutions for '*X with Y*' concepts in SNOMED CT for the frequently encountered clinical patterns in Table 3. The simplest modelling approach for representing X with Y concepts in SNOMED CT is:

1. Both *X* and *Y* are co-occurrent, but with no causality between *X* and *Y*.

X_{CLP} with Y_{CLP}, both simply asserted as co-occurring, and no known causal/manifestational relationship implied:

XwithY$_{\text{CLP}}$ equivalentTo X_{CLP} and Y_{CLP}

XwithY$_{\text{CLP}}$ denotes the class of life phases that are characterised by the full presence of both the conditions *X* and *Y*:

XwithY$_{\text{CLP}}$ equivalentTo '*Clinical life phase*' and '**has condition**' some *X* and '**has condition**' some *Y*

It can be shown that both definitions are equivalent, by logical transformation or by a reasoner like HermIT (2015). An important result (as represented by the second definition) is that for each time interval [t_1; t_2] any single human is considered to have one single life phase, which is characterised by the conditions that are wholly present in this interval. As discussed in Schulz *et al.* (2011b), the subsumption of complex disorder classes by their constituent disorder classes is a characteristic phenomenon in many disease / disorder terminologies, and the life phase interpretation puts it on a solid ground.

It can easily be shown that the same applies for complex clinical life phase types with more than two conjoints, e.g. in the case of the Tetralogy of Fallot (Schulz *et al.*, 2011b), a combined heart defect as an emblematic example.

2. *X* is due to *Y* but *X* and *Y* are not necessarily co-occurrent

Here, the correct way would be to assert causality between the conditions *X* and *Y*.

XcausedByY equivalentTo *X* and '**is caused by**' some *Y*

However, according to our interpretation, SNOMED CT disorder concepts are clinical life phases and the underlying conditions are not or are only indirectly available[2]. We therefore axiomatically extend the notion of causation and allow that a clinical life phase is causally related to another clinical life phase. To express this, we use the SNOMED CT relation '**due to**'.

> *XcausedByY*$_{CLP}$ equivalentTo *X*$_{CLP}$ and
> '**due to**' some *Y*$_{CLP}$

This is a simplification of the correct representation, which should be (using the BTL2 relation '**is caused by**'):

> **XcausedByY*$_{CLP}$ equivalentTo '*Clinical life phase*' and
> '**has condition**' some (*X* and '**is caused by**' some *Y*)

Due to the problem of referring to clinical conditions in SNOMED CT, in the modeling pattern we propose, '**due to**' connects two CLPs that are related by the fact that the first one has a condition that is caused by a condition that defines the second one. E.g., all instances of '*Disorder of optic chiasm due to non-pituitary neoplasm*'$_{CLP}$ imply an instance of '*Non-pituitary neoplasm*'$_{CLP}$ (which implies an instance of '*Non-pituitary neoplasm*'). We consider this approximation as sufficient for the reasoning services required.

3. *X* temporally follows *Y*. This does not specify that *X* is due to *Y*, although causality is frequently implied

> *XfollowsY*$_{CLP}$ equivalentTo *X*$_{CLP}$ and
> **follows** some *Y*$_{CLP}$

As mentioned before, the BTL2 relation **is preceded by** excludes the Allen relation **overlaps**. We argue that the relation **follows** should include **overlaps**, as it is common in medicine. E.g., *Postvaricella encephalitis* might include cases in which the varicella infection has not ended at the inception of the complication, viz. *Encephalitis*.

4. *X* is due to *Y* and both *X* and *Y* are co-occurrent
Here we propose a combination of patterns 1 and 2:

> *XdueToCooccurringY*$_{CLP}$ equivalentTo
> *X*$_{CLP}$ and *Y*$_{CLP}$ and '**due to**' some *Y*$_{CLP}$

[2] In the case of fully defined disorder concepts, the condition would correspond to the combination of location with morphology, inside the role groups

It looks uncommon that the same class *Y*$_{CLP}$ appears both as a superclass and a class related via the relation **due to**. Taking the example '*Hernia with intestinal obstruction*'$_{CLP}$: All life phases of this type are both *Hernia* life phases and *Intestinal obstruction* life phases, and they are related, additionally, to a second *Hernia* life phase (which is a different one but is assumed to refer to the same hernia object). This second life phase is, actually, one that precedes the inception of the complication, in this case the intestinal obstruction.

3.3 Special cases

There are other cases that we have excluded from our typology, but which, nevertheless, deserve consideration:

- Terms of the type '*Abscess of urethral gland due to Neisseria gonorrhoeae*'. Here, the right hand side of the particle "due to" denotes a material agent, not a process. In BTL2 this would be expressed by the relation '**has agent**' and could be modeled in the following way (note that in this case, the relation '**has condition**' is a paraphrase of the role group relation in SNOMED CT):

> *XwithAgentA*$_{CLP}$ equivalentTo *X*$_{CLP}$ and
> '**has condition**' some ('**has agent**' some *A*)

- Associativity. It can be shown that the following rule always holds and can be easily reduced to pattern one:

> *XwithY*$_{CLP}$ and *Z*$_{CLP}$ equivalentTo
> *X*$_{CLP}$ and *YwithZ*$_{CLP}$ equivalentTo
> *X*$_{CLP}$ and *Y*$_{CLP}$ and *Z*$_{CLP}$

E.g. '*Diabetes mellitus with hyperosmolar coma*'$_{CLP}$ superficially appears to be an *X* with *Y* pattern, but is more appropriately an example of *X* with *Y* with Z (*Diabetes mellitus*, *Hyperosmolar state*, *Coma*). There is no nesting.

- On this basis, more complex chained sequences according to pattern four are possible using the '**due to**' relationship. This might be represented as:

> '*Diabetes mellitus with hyperosmolar coma*'$_{CLP}$
> equivalentTo
> '*Diabetes mellitus*'$_{CLP}$ and
> ('*Hyperosmolar state*'$_{CLP}$ and
> '**due to**' some '*Diabetes*'$_{CLP}$) and
> (*Coma*$_{CLP}$ and
> '**due to**' some '*Hyperosmolar state*'$_{CLP}$)

4 DISCUSSION

Table 4 gives an overview of the relations used in the proposed models of our approach and their mapping to the Allen relations. As the BTL2 relation '**is preceded by**' does not allow overlap, it seems too strict. We prefer the relation **follows**, which makes the minimal assumption that the beginning of *Y* is later than X. This is also the assumption of the causality relation, which appears as a subrelation of **follows**. The proposal to include an overlap pattern for temporally following again blurs the distinction between pattern two and pattern three. This might be acceptable if we do not want to distinguish sequelae from other types of complications. When investigating definitions of sequelae under the concepts *Sequela* (finding) or *Sequela of disorders* (disorder), we found chronic or residual conditions that are complication of acute conditions that occur after the acute disease or injury phase. Sequelae can also be the result of the treatment of the primary condition. There is no time limit on when a late effect can occur; the residual condition may come directly after the disease or condition, or years later. This is a little vague in terms of whether the inciting condition is still present when the complication commences. In case there is a requirement to represent sequelae (late effects) as distinct from e.g., immediate complications, it might be worthwhile to define sequelae as not overlapping with their cause, and for this case to indeed use the BTL2 relation '**is preceded by**'.

Table 4. Allen relations compatible with the relations used in our models.

	Proposed Relations			
Allen Relations	Y co-occurs with X	Y is preceded by X	Y follows X	Y is due-to X
X takes place before Y		√	√	√
X meets Y		√	√	√
X overlaps with Y			√	√
Y starts X	√			
Y during X	√		√	√
Y finishes X	√		√	√
X is equal to Y	√			

5 CONCLUSION AND FURTHER WORK

The proposed patterns have been prototypically implemented in SNOMED CT and have achieved better semantic clarity and consistency in terminology creation and maintenance. The formal analysis of temporal and causative relationships has been proved to be useful for determining the patterns. Limitations identified and resulting tasks will be addressed by the ECE working group in the future:

- To evaluate if the proposed patterns are generic and can be applied throughout SNOMED CT, especially to concepts in the *Event* and *Procedure* hierarchies.
- To prove theoretically and empirically that the proposed patterns do not produce unexpected classification results, especially as a consequence of the simplification by asserting '**due to**' between situations and not between the underlying conditions.
- To check the impact of the new models on classification time.
- To extend the approach to negated conditions ("without") by scenarios that extend DL expressiveness or that represent negations as primitives. The impact on reasoning behaviour will also be investigated.
- To propose adjustments to the SNOMED CT naming conventions in the light of the new model.

REFERENCES

Allen, J. F (1983): Maintaining knowledge about temporal intervals. *Communications of the ACM* **26**(11), 832–843.

Baader, F. et al. (2007). The Description Logic Handbook. Cambridge: Cambridge University Press.

BFO (2015) Basic Formal Ontology. http://ifomis.uni-saarland.de/bfo/

HermiT (2015). HermiT OWL reasoner. http://www.hermit-reasoner.com/

IHTSDO (2010). Strategic Directions to 2015. http://www.ihtsdo.org/resource/resource/1

IHTSDO (2015) International Health Terminology Standards Development Organisation. SNOMED CT.

Motik, B. et al. (2012) OWL 2 Web Ontology Language Profiles (Second Edition). W3C Recommendation http://www.w3.org/TR/owl2-profiles/

NLM (2015). U.S. National Library of Medicine. Unified Medical Language System (UMLS), http://www.nlm.nih.gov/research/umls

Schulz, S. et al. (2011a). Scalable representations of diseases in biomedical ontologies. Journal of Biomedical Semantics. May **17**; 2 Suppl 2: S6.

Schulz, S. & Boeker, M. (2013). An Upper Level Ontology for the Life Sciences. Evolution, Design and Application. In: Furbach, U. & Staab, S. (eds.). Informatik 2013. IOS Press.

Schulz, S. et al. (2011b). Consolidating SNOMED CT's ontological commitment. Applied ontology **6**: 1-11.

Schulz, S. et al. (2012b). Competing interpretations of disorder codes in SNOMED CT and ICD. AMIA Annu Symp Proc. 819-827.

Smith, B. et al. (2005). Relations in biomedical ontologies. Genome Biology **6**(5): R46.

WHO (2015). International Classification of Diseases. http://www.who.int/classifications/icd/en/