0 EURECA ETL Guidelines

Table of contents

1 Express information with medical vocabularies

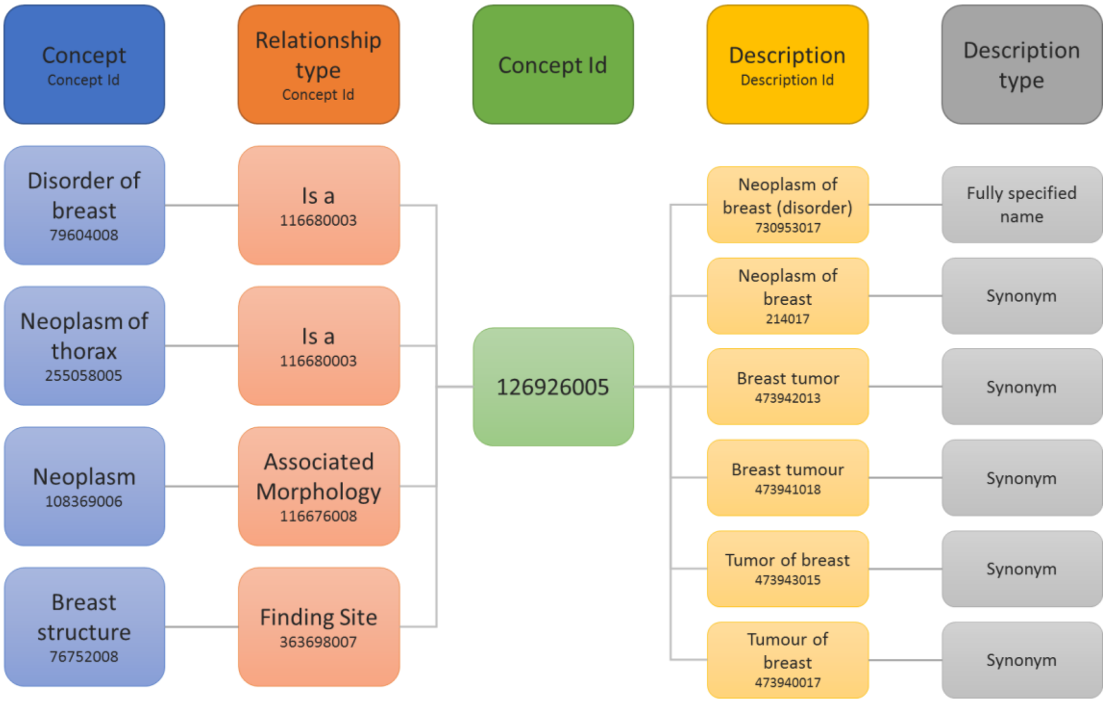
1.1 Using concepts from medical vocabularies to express information

1.1.1 SNOMED CT

SNOMED CT provides a standardized way to represent medical knowledge expressed by clinical phrases made by clinicians and enables automatic interpretation of these. It is a collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. The coverage of this terminology includes: : clinical findings, symptoms, diagnosis, procedures, body structures, organisms and other etiologies, substances, pharmaceuticals, devices and specimen.

This terminology contains concepts with unique meanings and formal logic based definitions organized into hierarchies, and its content is represented using three types of components:

- **Concepts**: Every concept has a unique numeric identifier and they represent clinical thoughts like "tumor" or "wound". Concepts are organized in a hierarchy, from more general to more detailed
- **Descriptions**: Link human readable terms to concepts. Each concept can have associated several descriptions. Each description represents a synonym that describes the same clinical concept, and each description has a unique numeric description identifier
- **Relationships**: Link concepts to other concepts which are related by semantic meaning in some way. They provide for definitions and other properties of the concept, and they are categorized in type of relationships and each relationship has a unique numeric identifier.



Apart from these three components, they are supplemented by "Reference Sets", which provide additional flexible features that enables customization and enhancement of the vocabulary. These includes language preferences or mapping from or to other code systems. Every reference set has a unique numeric concept identifier.

1.1.1.1 SNOMED CT hierarchy

SNOMED CT is structured following a polyhierarchy or a directed acyclic graph. This means that a given concept can be the descending of multiple parent concepts.

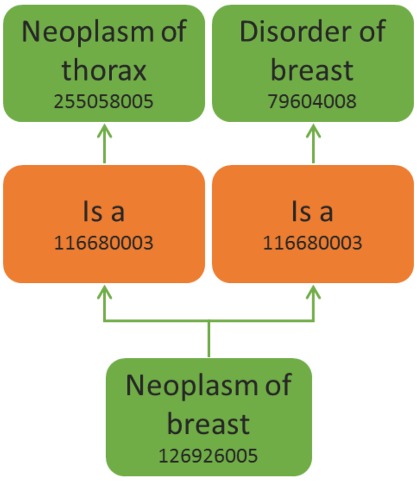


Figure 1: SNOMED CT Polyhierarchy

The image shows an example for the concept "Neoplasm of breast", which is the descendant of the concepts "Neoplasm of thorax" and "Disorder of breast" simultaneously.

Note that the polyhierarchy is defined using the "Is a" relationship type between concepts. This relationship is aimed uniquely to organize SNOMED CT in this special kind of hierarchy.

1.1.1.2 SNOMED CT attribute relationships

It has been appointed the "Is a" as the relationship type used to organized SNOMED CT as a polyhierarchy by relating two concepts. This relationship type because of its special purpose cannot be considered as the standard relationship type purpose, despite its function is to relate concepts.

In general, attribute relationships contribute to the definition of a concept, by associating it with the value of a defining characteristic. The main concept is related through a relationship type to another main concept, part of its definition, and resulting in a value-pair per each defining relationship (main concept -> relationship-type - value concept)

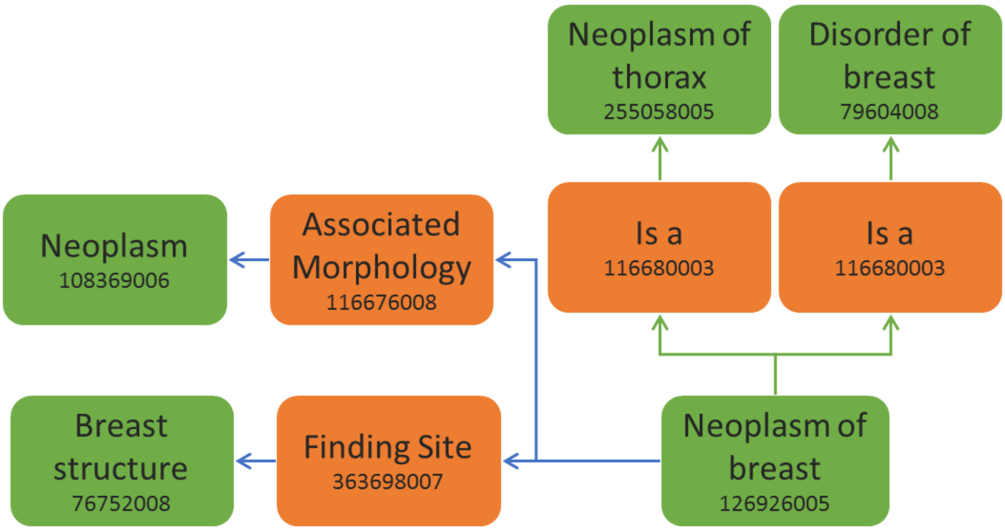


Figure 2: SNOMED CT concept relationships / defining characteristics

The image shows how the concept "Neoplasm of breast is related" to other four concepts by two "Is a" relationship types, and by two defining relationships "Finding site" and "Associated morphology". The defining relationships are composed by value-pairs, "Breast structure"-"Finding site", being the latter one the bridge or attribute concept that link the main concept "Neoplasm of breast" with the concept "Breast structure", related them. This value defines the concept and provide the meaning of where is the site of the observation "Neoplasm of breast". Therefore, the concept "Neoplasm of breast" is defined by two defining characteristics "Breast structure" and "Neoplasm", two concepts that together conforms the meaning of other concept.

Each relationship type has a domain and range constraints. The domain constraints limit the source concepts that the relationship links from.

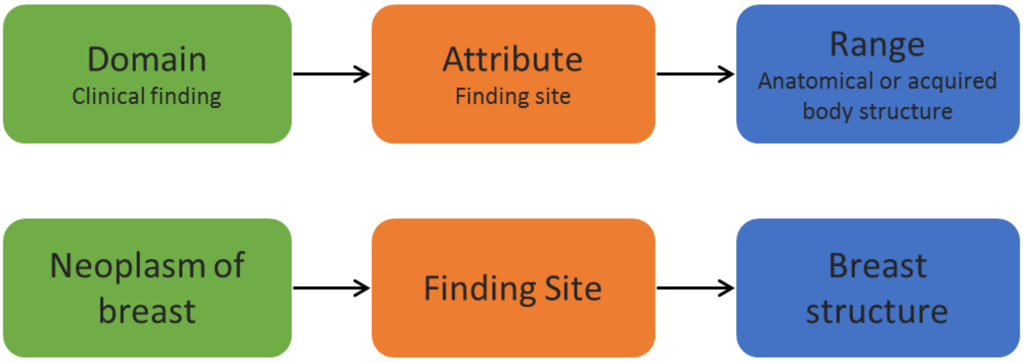


Figure 3: Attribute domain and range

From the example of the figure, the "Finding site" relationship type has a specific domain that is the hierarchy of concepts from "Clinical finding", being "Neoplasm of breast" the source concept part of this hierarchy. The range constraints limit the value concepts that the relationship links to. From the example of the figure, the "Finding site" relationship type has a specific range that is the hierarchy of concepts of "Anatomical or acquired body structure", being "Breast structure" the value concept and part of this hierarchy.

1.1.1.3 SNOMED CT normal form

Before describing the SNOMED CT normal form, it is required to define what is a fully-defined concept and what it is a primitive concept. A fully-defined concept has a set of defining characteristics or relationships to other concepts, which distinguish its meaning from other similar concept.

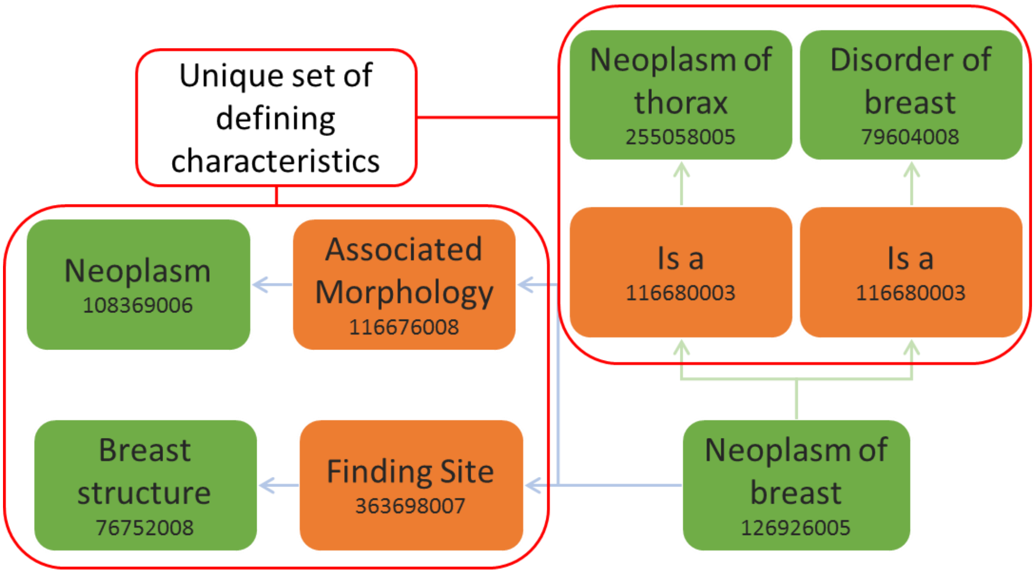


Figure 4: Unique set of defining characteristics

A unique set of defining characteristics implies that there are no other concept with the combination of relationships. In the figure, "Neoplasm of breast" is a fully defined concept, because there are no other concept in SNOMED CT with the following combination of relationships:

- “Is a” – “Neoplasm of thorax”
- “Is a” – “Disorder of breast”
- “Associated morphology” – “Neoplasm”
- “Finding site” – “Breast structure”

If the combination of defining characteristics is the same to the one of other concept, both concepts are primitive.

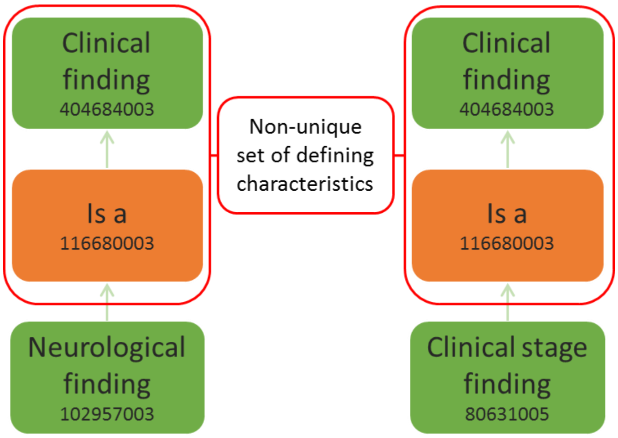


Figure 5: Non unique set of defining characteristics

The SNOMED CT normal form can be described as the decomposition of a concept into its defining characteristics that are not “Is a” relationships. Meaning that for a fully defined concept, its normal form it is the concept and its defining characteristics that are not “Is a” relationships. Obviously, for a primitive concept, its normal form it will be just the concept.

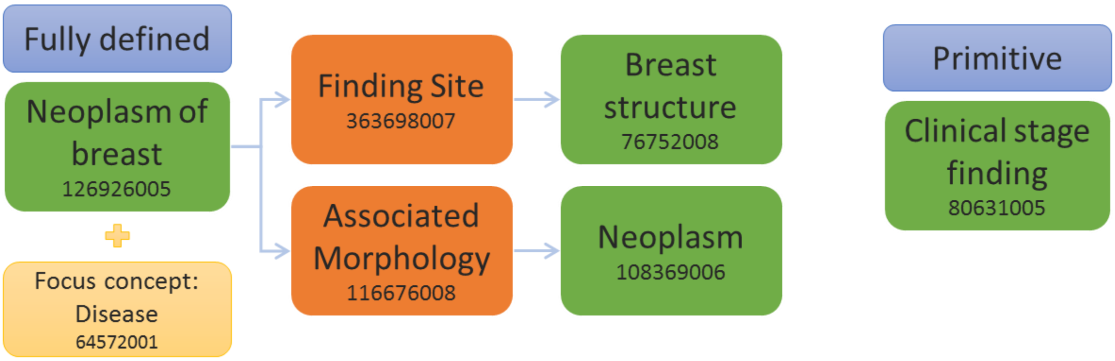

Fully defined
Neoplasm of breast 126926005
Focus concept: Disease 64572001
Finding Site 363698007
Breast structure 76752008
Associated Morphology 116676008
Neoplasm 108369006
Primitive
Clinical stage finding 80631005

Figure 6: SNOMED CT Normal form

In the figure, there is an addition as it could be noticed, which adds a focus concept "Disease" to the normal form of "Neoplasm of breast". The focus concept is the proximal primitive supertype [1]of the concept, which means that we must follow the ancestors in the hierarchy of the concept (going upwards in the hierarchy) until we find a primitive and supertype concept.

There are two different types of normal form, the long normal form and the short normal form. The normal form includes all the defining characteristics of the concept plus the focus concept, and on the other hand, the short normal form only includes the differential defining relationships of the concept.

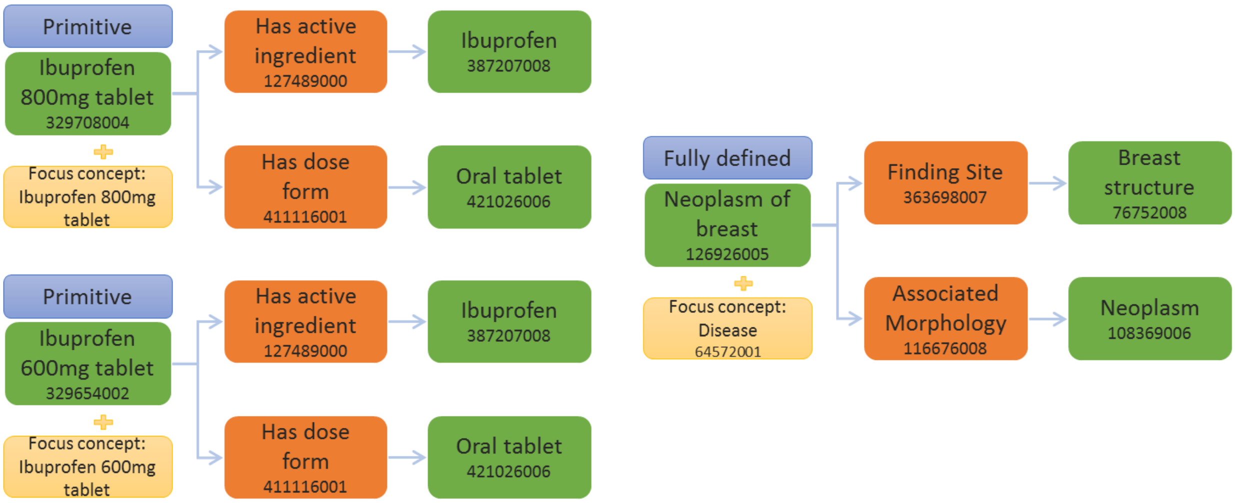

Primitive
Ibuprofen 800mg tablet 329708004
Focus concept: Ibuprofen 800mg tablet
Has active ingredient 127489000
Ibuprofen 387207008
Has dose form 411116001
Oral tablet 421026006
Primitive
Ibuprofen 600mg tablet 329654002
Focus concept: Ibuprofen 600mg tablet
Has active ingredient 127489000
Ibuprofen 387207008
Has dose form 411116001
Oral tablet 421026006
Fully defined
Neoplasm of breast 126926005
Focus concept: Disease 64572001
Finding Site 363698007
Breast structure 76752008
Associated Morphology 116676008
Neoplasm 108369006

Figure 7: SNOMED CT Long normal form

[1] Root concepts of SNOMED CT and direct descendants (Root concepts: Clinical finding, Procedure, Situation with explicit concept, Observable entity, Body structure, Organism, Substance, Pharmaceutical / biologic product, Specimen, Special concept, Physical object, Physical force, Event, Environments and geographical locations, Social context, Staging and scales, Qualifier value and Record artefact)

The figure shows an example of three long normal forms of two primitive concepts (which shares defining characteristics, and therefore they are primitive) and a fully defined concept. The long normal form as the figure shows, includes all the defining characteristics (or defining relationships) of the concept, despite they are unique or shared.

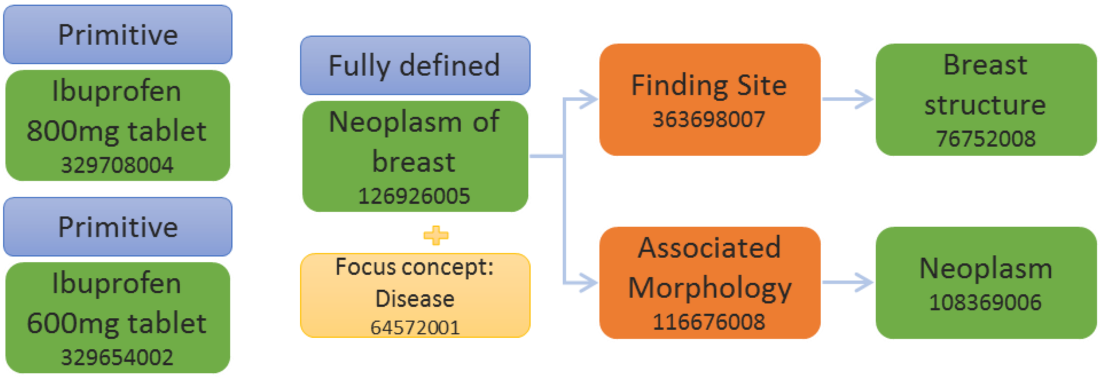

Primitive
Ibuprofen 800mg tablet 329708004
Primitive
Ibuprofen 600mg tablet 329654002
Fully defined
Neoplasm of breast 126926005
Focus concept: Disease 64572001
Finding Site 363698007
Breast structure 76752008
Associated Morphology 116676008
Neoplasm 108369006

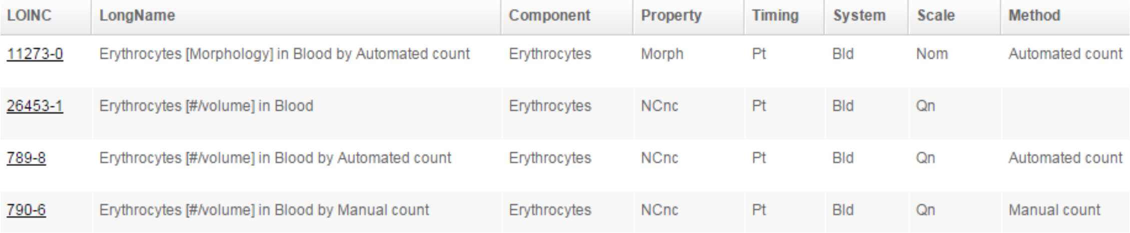
Figure 8: SNOMED CT Short normal form

The short normal form instead, includes the unique defining characteristics plus the focus concept. For this reason, as the figure shows, the short normal form for the primitive concepts, are just the concepts, and for the fully defined concept are the unique defining characteristics.

1.1.2 Other vocabularies

1.1.2.1 LOINC

LOINC is a vocabulary that identifies medical laboratory observations, being a domain specific terminology (differing of the general domain of SNOMED CT). It provides universal code names and identifiers to laboratory observations in order to assist in the electronic exchange and gathering of clinical results.



LOINC	LongName	Component	Property	Timing	System	Scale	Method
11273-0	Erythrocytes [Morphology] in Blood by Automated count	Erythrocytes	Morph	Pt	Bld	Nom	Automated count
26453-1	Erythrocytes [#/volume] in Blood	Erythrocytes	NCnc	Pt	Bld	Qn	
789-8	Erythrocytes [#/volume] in Blood by Automated count	Erythrocytes	NCnc	Pt	Bld	Qn	Automated count
790-6	Erythrocytes [#/volume] in Blood by Manual count	Erythrocytes	NCnc	Pt	Bld	Qn	Manual count

Figure 9: LOINC concepts and its six field specification

Despite the format of LOINC, with a 6-part name to include different fields for the unique specification of each identified single test, it doesn't represent a taxonomy or hierarchy type like SNOMED CT.

The possibility of including LOINC as part of a general taxonomy for a common core dataset, which can be composed by multiple medical vocabularies, can be achieved. A possible approach is to due to the domain specific of the LOINC terminology, is that it can be part of a more general taxonomy that includes observations, and therefore include all LOINC concepts under this classification.

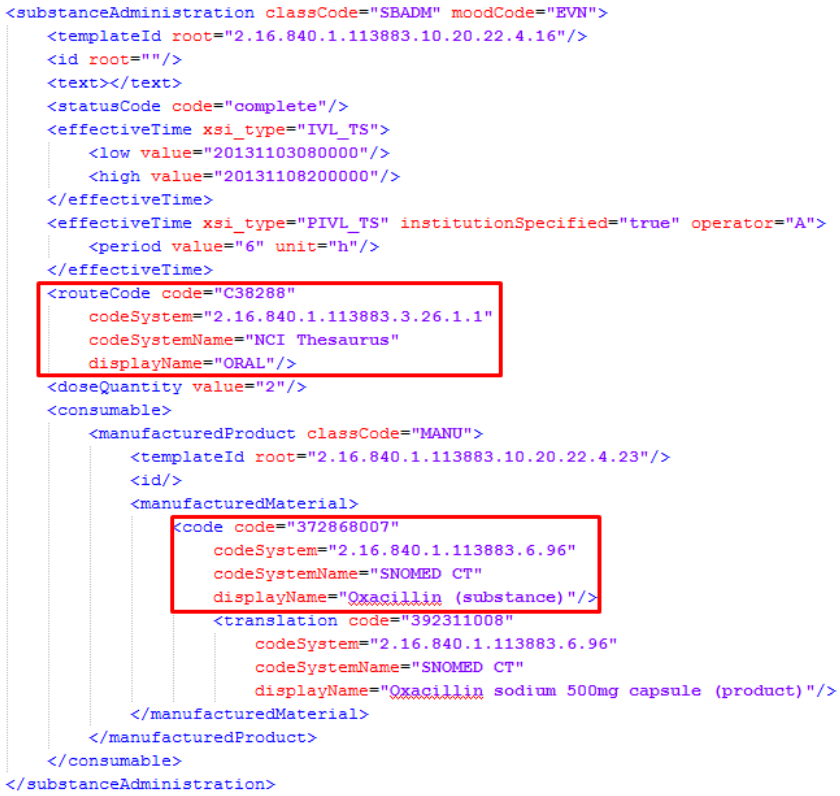
1.1.2.2 HGNC

The purpose of this vocabulary is to approve a unique and meaningful name for every known human gene based on a query of experts, being as LOINC, a domain specific vocabulary. The committee have established a naming guidelines to assign a symbol and name to each gene, as well as a naming procedure to contact authors who have published the human gene in order to coordinate the inclusion of a new gene.

As a domain specific vocabulary, the same approach used for the LOINC terminology can be applied to HGNC in order to be included into a more generalistic taxonomy. For example, the HGNC terms could be covered by an Entity/gene hierarchy of other more general terminology or group of terminologies.

1.2 Composition of concepts

In order to express more detailed information, that cannot be represented or isn't available by a single concept, more than one concept may be used. In fact, it is common the use of different concepts into a CDA document, in order express complex information with the use of multiple concepts combined. For example, in order to express a "Medication activity" by a CDA document, it is required the use of different concept from different terminologies to express the route of administration (NCI Thesaurus), the body site (SNOMED CT) and the product (SNOMED CT).



```
<substanceAdministration classCode="SBADM" moodCode="EVN">
    <templateId root="2.16.840.1.113883.10.20.22.4.16"/>
    <id root=""/>
    <text></text>
    <statusCode code="complete"/>
    <effectiveTime xsi_type="IVL_TS">
        <low value="20131103080000"/>
        <high value="20131108200000"/>
    </effectiveTime>
    <effectiveTime xsi_type="PIVL_TS" institutionSpecified="true" operator="A">
        <period value="6" unit="h"/>
    </effectiveTime>
    <routeCode code="C38288"
        codeSystem="2.16.840.1.113883.3.26.1.1"
        codeSystemName="NCI Thesaurus"
        displayName="ORAL"/>
    <doseQuantity value="2"/>
    <consumable>
        <manufacturedProduct classCode="MANU">
            <templateId root="2.16.840.1.113883.10.20.22.4.23"/>
            <id/>
            <manufacturedMaterial>
                <code code="372868007"
                    codeSystem="2.16.840.1.113883.6.96"
                    codeSystemName="SNOMED CT"
                    displayName="Oxacillin (substance)"/>
                    <translation code="392311008"
                        codeSystem="2.16.840.1.113883.6.96"
                        codeSystemName="SNOMED CT"
                        displayName="Oxacillin sodium 500mg capsule (product)"/>
            </manufacturedMaterial>
        </manufacturedProduct>
    </consumable>
</substanceAdministration>
```

Figure 10: Medication CDA document using multiple concepts

What has been presented could raise doubts because in fact a CDA document is a way to structure a lot of information, and therefore the use of multiple concept is required to express structured information. But the main purpose is to provide a different approach, in order to be able to differentiate the use of multiple concepts to represent structured information, and the use of them to express specific information.

Vocabularies as SNOMED CT provides mechanisms to create a composition of concepts that can express a clinical phrase. This SNOMED CT mechanisms is denominated post coordination, and follows a set of rules that mimic the relationships value-pair, that are the defining characteristics of SNOMED CT concepts.

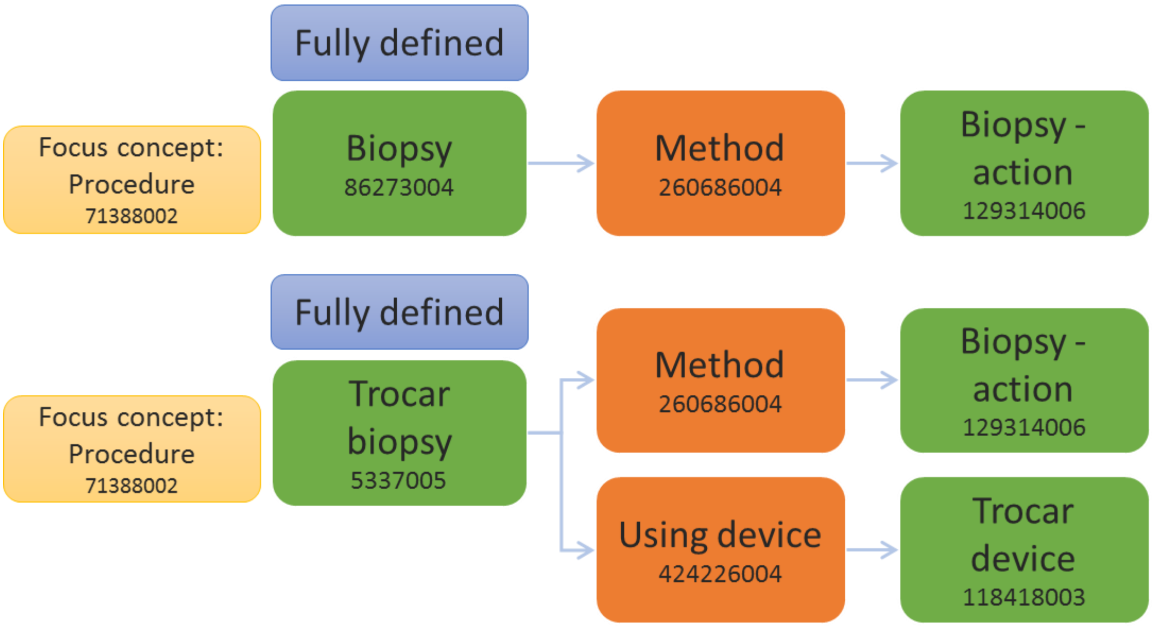


Figure 11: SNOMED CT short normal form of similar concepts

Using the example on the figure, we can compose a postcoordinated expression (composition of concepts in SNOMED CT) despite there already is a single concept to express the target information in SNOMED CT. If we want to express the information relative to a biopsy of a patient where a trocar has been used, we can compose this information by:

- Biopsy 86273004
    - Using device 42422600 = Trocar device 118418003

This post coordinated expression is coded as 86273004|Biopsy|:42422600|Using device|=118418003, and has exactly the same meaning than the single concept “Trocar biopsy 5337005”. These is because the defining characteristics of the concept "Biopsy 86273004" plus the addition of "Using device - Trocar device", have the same defining characteristics than the concept "Trocar biopsy 5337005".

If we wanted to make a composition of concepts to express clinical information that cannot be express by a single SNOMED CT concept, for example, "Trocar biopsy of the surface of the rib". Specifically at this case there is no such a concept in SNOMED CT, and therefore, if we want to express this information we have to use postcoordination:

- Biopsy 86273004
    - Using device 42422600 = Trocar device 118418003
    - Procedure site 363704007 = Surface of rib 360946008

By this composition we have achieved to express the intended meaning, of a trocar biopsy at the surface of the rib bone. But there could be an alternative way of compose this meaning:

- Trocar biopsy 5337005
    - Procedure site 363704007 = Surface of rib 360946008

Both composition contains the same defining characteristics, from the main concept, or artificially added, and therefore they express the exact same meaning.

This mechanisms is a power resource provided by SNOMED CT that enables the expression of any kind of information due to its flexibility. Additionally, because the rules of the defining characteristics and the normal form are followed, this process can work seamless with the normal form, so they are structured alike.

1.3 Modelling information based on HL7 RIM Acts

In order to interoperate between clinical information, expressed with medical terminologies and a system that make use of HL7 v3 standards, it is required to follow a general approach to successfully achieve this interoperability.

The focus of health care communication and documentation using HL7 standards are the actions taken to treat a patient:

- Request or order for a test is an action
- Report of the rest result is an action
- Creating a diagnosis based on test results is an action
- Prescribing treatment based on the diagnosis is an action

A medical record is a record of each of the individual actions that make up the diagnosis, treatment and care of a patient. In every action or act there are a set of entities that take part in the action, and to represent this kind of information, HL7 has established five core concepts:

- Every happening is an **'Act'** (Procedures, observations, medications, supply, registration,etc)
- Acts are related through **'ActRelationships'** (Composition, preconditions, revisions, support,etc)
- **'Participation'** defines the context for an Act (Author, performer, subject, location, etc)
- The participants are **'Roles'** (patient, provider, practitioner, specimen, employee, etc)
- Roles are played by **'Entities'** (Persons, organizations, material, places, devices, etc)

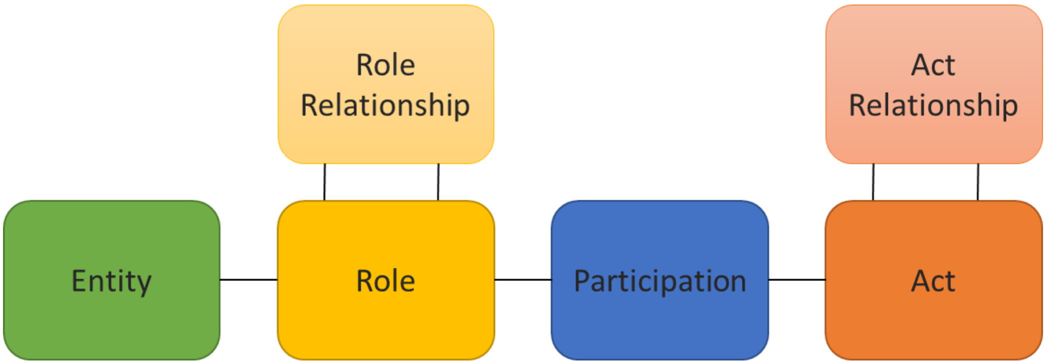


Figure 12: HL7 RIM representation through actions

Due to this approach followed by HL7 standards, it is required to model clinical information according to this pattern. The purpose of this section is to provide some examples in the clinical trial cancer domain. It is going to be provided guidance for some troublesome representations of information using the action approach, including some particular cases.

1.3.1 Associated morphologies

At the cancer domain is common the need to express information about diseases. In many cases the disease changes the tissue or cellular level morphology, being this change a characteristic feature of the disease.

In order to express this information with SNOMED CT, this terminology provides a hierarchy to represent information about the characteristic changes (tissue or cellular level) of a disease. The concept "49755003 Morphologically abnormal structure" and its subtype descendants are the concepts used to represent this kind of information.

Despite that SNOMED CT has a hierarchy for this kind of information, it is common the use of post-coordinate concepts, which contain this information with one defining relationship, through the "116676008 Associated morphology" relationship type. The main reason for this, is because they are concept already included at the SNOMED CT that are commonly defined for two defining relationships, "116676008 Associated morphology" and "363698007 Finding site", being the last one use to define the specific characteristic about the body site affected by the condition.

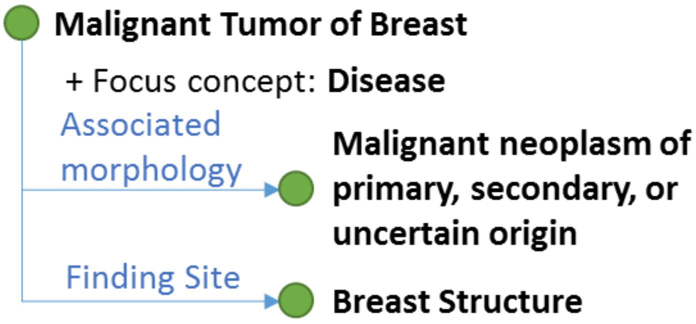

Malignant Tumor of Breast
+ Focus concept: Disease
Associated morphology
Malignant neoplasm of primary, secondary, or uncertain origin
Finding Site
Breast Structure

Figure 13: Defining characteristics of a SNOMED CT concept

Sometimes, due to the peculiarity of the information to be expressed, SNOMED CT doesn't include the required concept that specifically links an associated morphology and a body site. Therefore it is require to make a composition of concepts to express this information, joining a concept from the "49755003 Morphologically abnormal structure" subtype ("116676008 Associated morphology" range) with a concept from the "442083009 Anatomical or acquired body structure" ("363698007 Finding site" range).

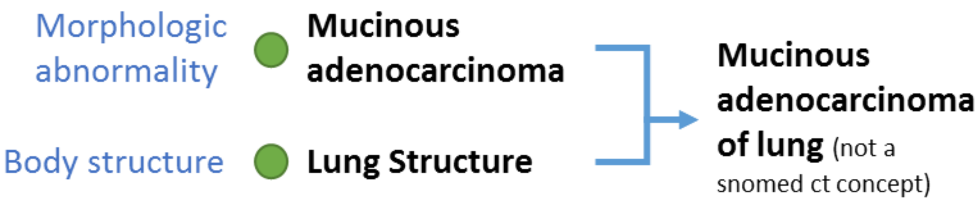
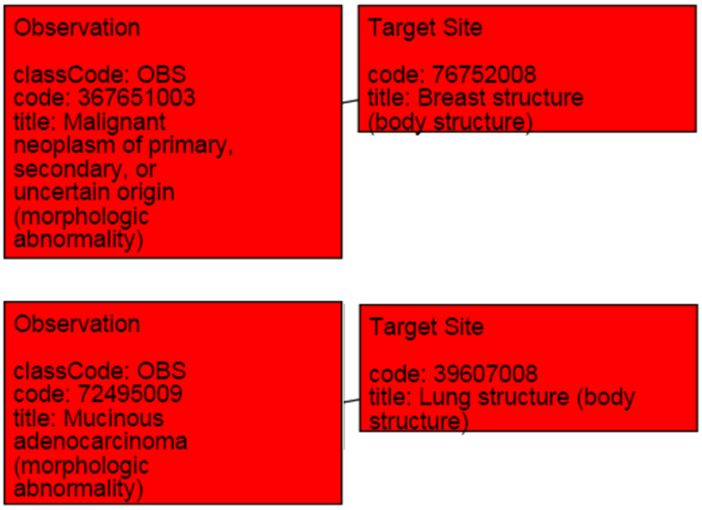


Figure 14: Composition of concepts to express a "Mucinous adenocarcinoma of lung"

This situation presents an issue when trying to be modelled within a HL7 RIM based data warehouse. As it has been commented, we should try to represent this information as acts, specifically for this case as observation acts. When try to apply this approach when considering whether a morphologic abnormality is an act, we have found a void in the semantic interoperable guidelines between SNOMED CT and HL7 RIM based models.

Therefore, we are currently accepting morphologic abnormalities as acts of observations within this paradigm. Additionally, the morphologic abnormality from the defining relationships of a SNOMED CT concept are currently considered the main concept of the act. This situation entails that the observation act code will be the associated morphology, and it will be related with the required additional information of observation acts, like the body site affected.


Observation

classCode: OBS
code: 367651003
title: Malignant neoplasm of primary, secondary, or uncertain origin (morphologic abnormality)

Target Site

code: 76752008
title: Breast structure (body structure)

Observation

classCode: OBS
code: 72495009
title: Mucinous adenocarcinoma (morphologic abnormality)

Target Site

code: 39607008
title: Lung structure (body structure)

Figure 15: Morphologic abnormalities observation acts. HL7 RIM based representation

As the figure shows, we have the two examples provided along this section. The first one is the "Malignant tumor of breast", which is decomposed into its defining relationships. As it can be observed, the code of the observation act is not the code for

"Malignant tumor of breast" being the code of the associated morphology "Malignant neoplasm of primary, secondary, or uncertain origin" instead. This morphologic abnormality is complemented with the attribute "Target site" with the code of the finding site of the other defining relationship from the decomposition of the original concept.

The second example shows how the manual composition of concept to express a "Mucinous adenocarcinoma of lung" is represented within the HL7 RIM based data model. The morphologic abnormality is used directly as code for the observation act, and then complemented with the code of the finding site of the "Lung structure".

1.3.2 Administration of substances

There is a considerable amount of information involved with medication activities that can be collected, processed and stored using a framework based on HL7 and the support of medical vocabularies like SNOMED-CT.

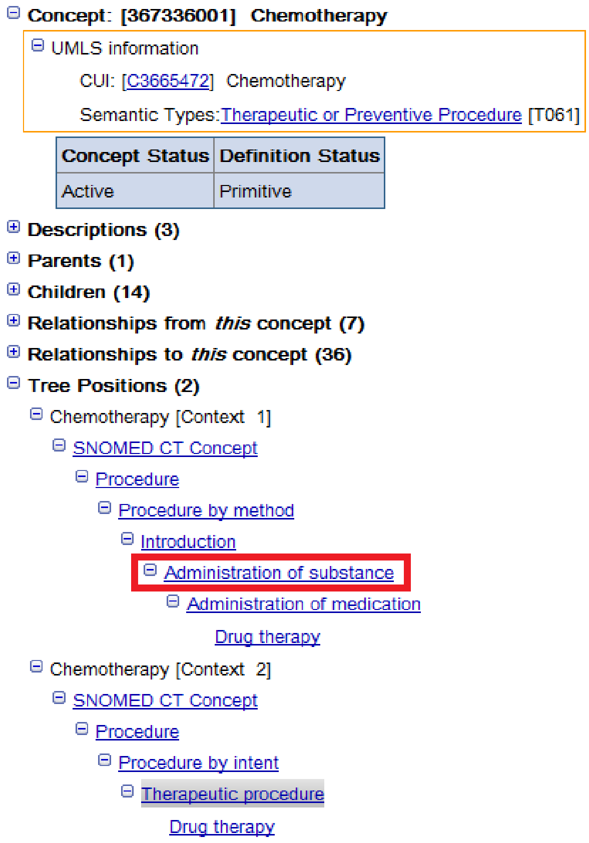
In the perspective of the data model all this information is organized around an Act of SubstanceAdministration class (medical event representing a drug therapy), with a participating Entity which corresponds to the administered product or substance. Attributes of the Act also store specific information like the administered dose or the date of the therapy.

The distinctive element between SubstanceAdministration acts is the substance/product, as the act cannot include more than one participating product. So if a therapy is composed of several products, it would require one act for each, as a medical prescription commonly does. Note that a given product can be composed of multiple substances or active ingredients.

It is necessary to map a portion of this information to concepts of medical terminologies, but not all of the information requires a mapping. Because some information could be modelled using the HL7 predefined structures of the messages, and then inserted directly in the CDM.

Below is a description of all the data related with substance administration therapies that can be stored, with an indication of whether they require a concept to be modelled or not:

- Type of procedure performed: This is the main element of the Act, the rest of the elements are characteristics applied to this element. It is mapped with SNOMED-CT concepts organized under the ‘Administration of Substance’ branch, which is classified in the ‘Procedure’ semantic type of SNOMED CT. In SNOMED CT each branch identifier is also a concept, so the ‘Administration of Substance’ branch is identified by a concept with that label and ‘432102000’ code. This concept is the default value of this component, but it could also be any other son of that concept, for example ‘Chemotherapy’ (367336001) or ‘Pre-operative chemotherapy’ (394894008).



⊟ **Concept: [367336001] Chemotherapy**

⊟ UMLS information

CUI: [C3665472] Chemotherapy

Semantic Types: Therapeutic or Preventive Procedure [T061]

Concept Status	Definition Status
Active	Primitive

⊞ **Descriptions (3)**

⊞ **Parents (1)**

⊞ **Children (14)**

⊞ **Relationships from *this* concept (7)**

⊞ **Relationships to *this* concept (36)**

⊟ **Tree Positions (2)**

- ⊟ Chemotherapy [Context 1]
  - ⊟ SNOMED CT Concept
    - ⊟ Procedure
      - ⊟ Procedure by method
        - ⊟ Introduction
          - ⊟ Administration of substance
            - ⊟ Administration of medication
              - Drug therapy
- ⊟ Chemotherapy [Context 2]
  - ⊟ SNOMED CT Concept
    - ⊟ Procedure
      - ⊟ Procedure by intent
        - ⊟ Therapeutic procedure
          - Drug therapy

Figure 16: Display of the tree position of 'Chemotherapy' concept in SNOMED-CT branches using UMLS Terminology Services

- Method of administration used: This information indicates the procedure performed to apply the medication dose. It should also be represented with an SNOMED CT concept for example 'Injection – action' (129326001) of 'Qualifier Value' semantic type.

- Substance or product administered: Represents the Entity participating in the Act. It should be mapped with concepts of the SNOMED CT 'Pharmaceutical/biologic product' or 'Substance' semantic types, for example 'Hydroxycarbamide' (387314007).

- Body structure where the method is applied: Represents the target site of the Act. It should be mapped with SNOMED CT concepts of 'Body structure' semantic type for example 'Venous structure' (29092000).

- The way or router that the substance follows when it is administered: In the HL7 RIM it is referred to as 'routeCode', an attribute of the SubstanceAdministration class. There is a branch in SNOMED CT suitable for these concepts called 'Route of administration value', of 'Qualifier Value' semantic type. For example 'Intravenous route' (47625008). But is required to use the HL7 RouteOfAdministration vocabulary NCI Thesaurus (e.g. 'INTRAVENOUS', 'ORAL', 'RECTAL', etc) for messaging purposes.

- The description of the body site that is used to approach to the real site of the administration is provided by the approach site code. It should be mapped with SNOMED CT concepts.

- Trigger/Reason for the prescription of the medication: It could also be stored the observation that led to the indication of the therapy modelling it as a related Act. For example SNOMED CT concepts of the 'Disorder' semantic type like a 'Carcinoma of breast' (254838004) that should be treated with chemotherapy.

For the following data there are structures defined by HL7 that collect them precisely inside of the CDA messages and subsequently into the data model. Therefore these data do not require any annotation into another terminology, because its variability (e.g. different numerical values) can be collected without using an external medical terminology.

- Conception of the act: It is defined as an attribute for Acts in the HL7 RIM called 'moodCode'. According to the HL7 official wiki site, "*The Act.moodCode is a structural code which is defined as a code distinguishing whether an Act is conceived of as a factual statement or in some other manner as a command, possibility, goal, etc. Its values are drawn from the HL7 ActMood vocabulary table.*" There are many possible values defined for this attribute, but the ones most commonly used in substance administration contexts are 'PRP' (proposal) and 'EVN' (event).

- Status of the act: To specify if a medication therapy is planned, active, completed, postponed, abandoned, etc…it is partially equivalent to the moodCode.

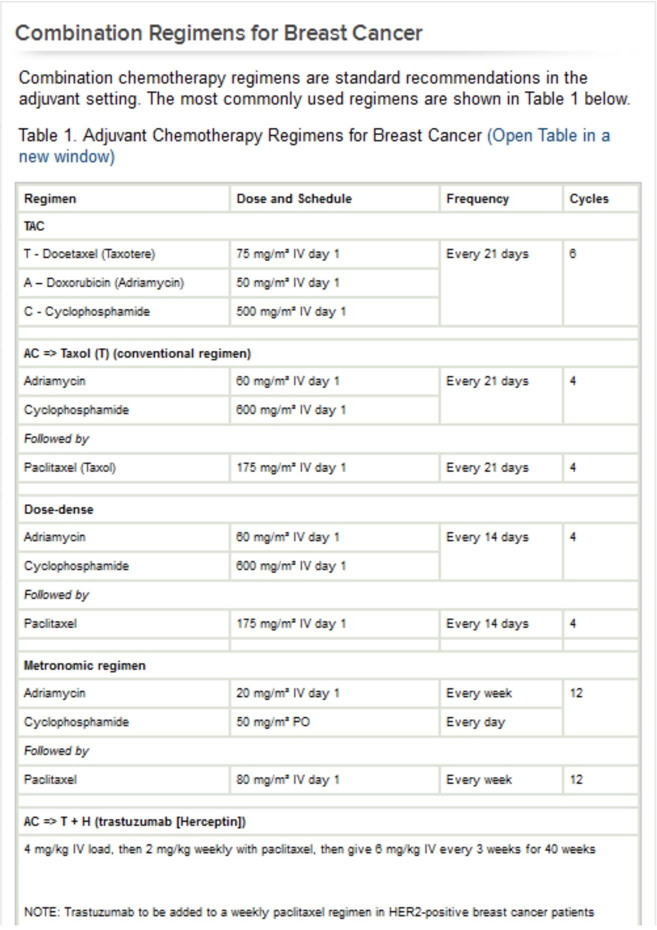
- Medication dose and its unit: Both planned and real doses (for finished events) of the therapy can be stored along with their units in UCUM standard format.

- Frequency of administration of a dose and its time unit: The time intervals between the administrations of each dose of medication along with their units in UCUM standard format.

- Rate of administered substance per time unit: The rate is given in units that have measure over time (e.g. 150mg for 1 hour)

- Start and end dates of the therapy: It could be implicit in the prescription and may be calculated using the start date and the number of medication cycles and their frequency.

Combining all these elements, complex medication therapies as the ones showed in Figure 2 can be modelled without any loss of information.



Combination Regimens for Breast Cancer

Combination chemotherapy regimens are standard recommendations in the adjuvant setting. The most commonly used regimens are shown in Table 1 below.

Table 1. Adjuvant Chemotherapy Regimens for Breast Cancer (Open Table in a new window)

Regimen	Dose and Schedule	Frequency	Cycles
TAC			
T - Docetaxel (Taxotere)	75 mg/m^2 IV day 1	Every 21 days	6
A – Doxorubicin (Adriamycin)	50 mg/m^2 IV day 1		
C - Cyclophosphamide	500 mg/m^2 IV day 1		
AC => Taxol (T) (conventional regimen)			
Adriamycin	60 mg/m^2 IV day 1	Every 21 days	4
Cyclophosphamide	600 mg/m^2 IV day 1		
Followed by			
Paclitaxel (Taxol)	175 mg/m^2 IV day 1	Every 21 days	4
Dose-dense			
Adriamycin	60 mg/m^2 IV day 1	Every 14 days	4
Cyclophosphamide	600 mg/m^2 IV day 1		
Followed by			
Paclitaxel	175 mg/m^2 IV day 1	Every 14 days	4
Metronomic regimen			
Adriamycin	20 mg/m^2 IV day 1	Every week	12
Cyclophosphamide	50 mg/m^2 PO	Every day	
Followed by			
Paclitaxel	80 mg/m^2 IV day 1	Every week	12
AC => T + H (trastuzumab [Herceptin])			
4 mg/kg IV load, then 2 mg/kg weekly with paclitaxel, then give 6 mg/kg IV every 3 weeks for 40 weeks NOTE: Trastuzumab to be added to a weekly paclitaxel regimen in HER2-positive breast cancer patients			

Figure 2: Adjuvant chemotherapy regimens for breast cancer, extracted from Medscape, as an example of a complete treatment that could be modelled with our framework.

1.3.3 Genetic findings

Genetic test results and all their details can be modelled completely with the combination of the HL7 framework and medical terminologies.

These medical events consist on the analysis of the status of a gene or protein in a patient, to support his diagnosis, to predict the response of the patient to a concrete therapy or to determine his prognosis.

This set of information is also organized under an Act, of Observation class, and its associated parameters. One element to remark is the participating Entity of the Act. To represent this information HGNC vocabulary (HUGO Gene Nomenclature Committee[2]) will be used. Given the short coverage of genes and proteins in SNOMED-CT, HGNC is preferred, to cover all the possibilities in genetic tests. Nevertheless, SNOMED-CT includes some genetic entities in its ‘Substance’ branch.

The HUGO Gene Nomenclature Committee has established a standard nomenclature for human genes. It consists in a short-form abbreviation known as gene symbol (for example ‘ERBB2’), and a longer and more descriptive name (for example ‘v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2’). This standard ensures that each symbol is unique and each gene is only represented by one approved gene symbol.

[2] http://www.genenames.org/

Below is a description of all the elements of these observations that can be collected, describing if a mapping to a concept of a medical terminology is needed or not:

- Evaluation procedure: Main element of the Act that establishes the whole representative process of this medical event. It gathers the applied method, the data obtained, its interpretation and classification, etc…being all these elements (described below) characteristics applied to this Evaluation Procedure. It requires an SNOMED-CT concept for its representation, which will be the main code of the Act. To select a concept that fulfills this requirement of representing the process in a global way, generic and wide concepts will be selected.
  Some suitable concepts for this element would be:
  - 'Genetic test' (405824009)
  - 'Molecular genetic test' (405825005)
  - 'Genetic finding' (106221001)

  The concept could also join some of the elements that are going to be described later in one post-coordinated concept. For example 'Human epidermal growth factor receptor 2 gene detection by fluorescence in situ hybridization' (434363004). It includes the genetic entity ('Human epidermal growth factor receptor 2'), the evaluation procedure ('gene detection') and the method performed ('fluorescence in situ hybridization').

- Measured entity: Protein or gene measured in the test. This entity is the subject of the observation. To be represented in a homogenous way according to the data stored in the EURECA CDM, a HGNC concept should be selected to map this element. As it has been mentioned before, HGNC also provides more variety in genetic entities than SNOMED-CT.
  Some examples would be:
  - MKI67 ('marker of proliferation Ki-67')
  - ERBB2 ('v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2')
  - PGR ('progesterone receptor')
  - ESR1 ('estrogen receptor 1')

- Status of the act: Defined in 2.1.2 section, does not need any mapping. For this cases the status should be 'completed', otherwise results could not be registered (scores, interpretations).

- Date of performance of the test: Effective time when the test took place (start and end dates can be registered with hours, minutes and seconds). Does not need to be mapped with a concept.

- Observation values: This element contains the result of the test, as a numerical score or percentage representing for example the amount of protein present on the cells surface of a tissue sample. It does not require an associated concept since it is stored as a numerical value. One example of these values would be the results obtained after performing a test for detection of protein levels (e.g. hormone receptors) by immunohistochemistry (IHC). This test consists on a staining process that makes the hormone receptors show up in a sample of breast cancer tissue. The values obtained after the performance of the test are a proportion score (percentage of positive cells, or its translation to a standard score from 0 to 8, where for example value 3 represents from 1/10 to 1/3 of colored cells) and an intensity score (a scale of values representing the intensity of the staining, e.g. value 2 represents moderate staining). The exact process is described in Harvey et al. J. Clin. Oncol 1999; 17:1474-1481[3].

- Reference range of values: Describes the variations of a measurement, delimiting the lowest and highest possible values, to provide guide for interpretation.

- Interpretation: Codes that specify interpretation of genetic analysis, such as "positive", "negative", "carrier", "responsive", etc. All the possible values are collected in the HL7 ObservationInterpretation value set, so this element does not need a mapping concept. In some cases the threshold between negative and positive interpretation values depends on an agreed percentage of gene presence in the cells. For example in the case of assessing the estrogen receptor status by immunohistochemistry, an IHC score of greater than 2 (corresponding to as few as 1% to 10% weakly positive cells) was used to define ER positivity. The interpretation code can be set to negative for score <= 2. For progesterone receptor, on the other hand, there is no commonly agreed cut-off yet, so no interpretation code can be set.

- Measurement method: Procedure performed to test gene levels. It needs an SNOMED-CT mapping, for example ‘Fluorescence in situ hybridization’ (426329006) or ‘Immunohistochemistry procedure’ (117617002) (both from HL7 receptor status measurement methods value set). Both are classified in ‘Laboratory Procedure’ branch in SNOMED-CT.

1.3.4 Scorings (severities)

Observations about the adverse effects of drugs are pretty common in the domain of clinical trial for cancer, as part of cancer therapy. These observation tries to gather and to report undesired harmful effects resulting on the patient from a medication of a cancer therapy.

The severity of the side effect is scored with the "Common Terminology Criteria for Adverse Events (CTCAE)" or "Common Toxicity Criteria (CTC)", which are a set of criteria for the standardized classification of adverse effects of drugs used in cancer therapy. CTC is a product of the US National Cancer Institute (NCI), and it is the most used system in US and UK drug trials. The observations has a range of grades from 1 to 5, depending on the severity.

1. Mild
2. Moderate
3. Severe
4. Life threatening
5. Death

In order to express this information, is therefore required to express the information relative to the observations itself, as well as the result of the severity of this observation.

[3] http://jco.ascopubs.org/content/17/5/1474

The observation will be an act of observation, expressed with a SNOMED CT concept (e.g. "422587007 Nausea), and for the severity SNOMED CT is used as well. For expressing severities, SNOMED CT provides a set of concepts under the hierarchy of "272141005 Severities":

- Severities 272141005
  - Mild 255604002
  - Mild to moderate 371923003
  - Moderate 6736007
  - Moderate to severe 371924009
  - Severe 24484000
  - Life threatening severity 442452003
  - Fatal 399166001

Through this set of SNOMED CT concepts the severity range of the CTC can be covered.

Mild	Mild 255604002
Moderate	Moderate 6736007
Severe	Severe 24484000
Life threatening	Life threatening severity 442452003
Death	Fatal 399166001

In order to pair this two concepts together into the an HL7 RIM data model, the concept of the observation of the side effect will be use as the code for the act of observation, and the concept from the severity hierarchy will be used as value of the observation.

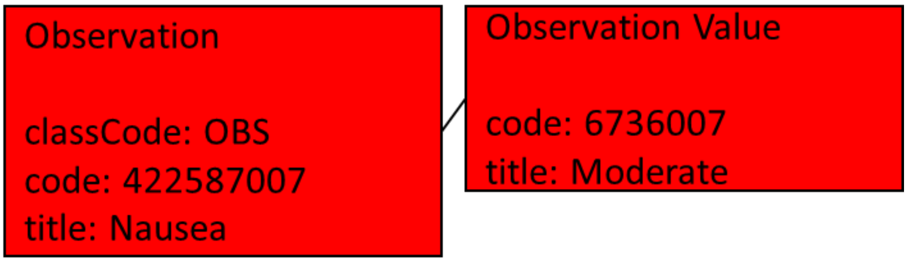


Figure 17: Side effect severity representation in a HL7 RIM based model

As the figure shows, the observation of the side effect "Nausea" is paired with the severity of the observation as an observation value, in this case "Moderate".

1.3.5 Breast Cancer Diagnosis

In order to express a diagnosis, which is the determination of which disease or condition is causing a person's signs and symptoms, we must include all the supporting observations that sustain the diagnosis. All the supporting observations, along with other information like the type of tumor, the exact site of the tumor, the staging and post treatment status if possible, must be structured specifically so it can contain all the required information for a diagnosis determination.

All the information that must be included for the determination of a "Breast Cancer Diagnosis" must be expressed like act or a set of them related by specific relationships. The purpose of this is to provide an overview of this approach that will ease the process of HL7 message creation to make this information interoperable.

For the "Breast Cancer Diagnosis", we will required the use of a set of acts. One of them will be the main act, and it will be related to other acts that will provide additional information. The main act is the act of the diagnosis with the kind of tumor, and the site of it. This act will be related to other acts that will support the diagnosis (staging and post treatment status).

The main act contains the statement of the diagnosis, as well as information about the type of cancer and its site. In order to express this information it is required the code for "Diagnosis 282291009" paired with a value that will be the code for the kind of cancer (e.g. Infiltrating duct carcinoma of female breast 448952004). The possible values for the diagnosis must be in the range of the value set “Breast Cancer 2.16.840.1.113883.3.526.3.389”. This main act requires a site, where the cancer is present, and must be expressed with a code in the range of the value set "Body Site Value Set 2.16.840.1.113883.3.88.12.3221.8.9", mainly represented by the SNOMED CT terminology (e.g. “Left breast 80248007”).

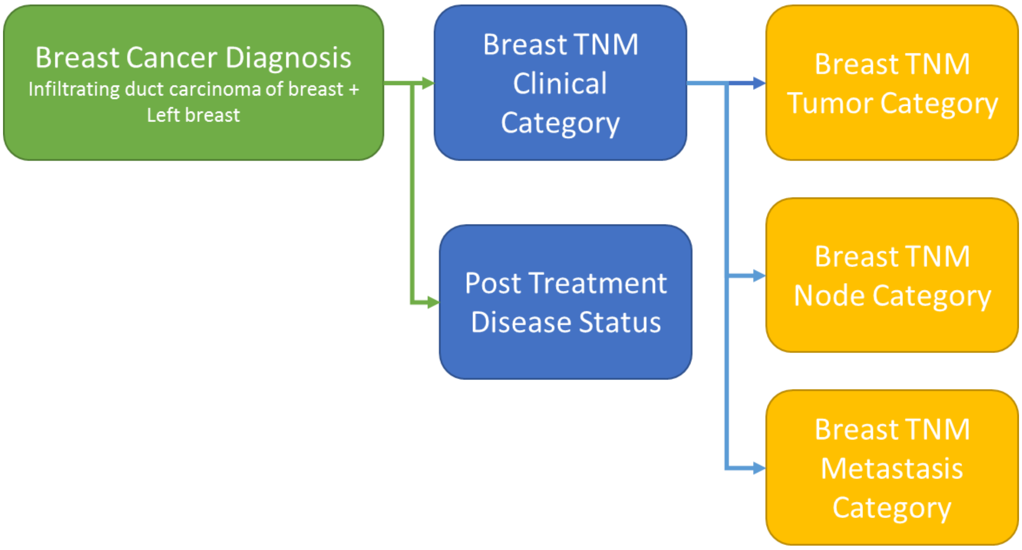

Breast Cancer Diagnosis
Infiltrating duct carcinoma of breast + Left breast
Breast TNM Clinical Category
Post Treatment Disease Status
Breast TNM Tumor Category
Breast TNM Node Category
Breast TNM Metastasis Category

Figure 18: Cancer diagnosis supporting data

Once defined the fundamental information relative to the diagnosis of breast cancer, it is required additional data that supports and complement the diagnostic. Note that the staging information is required because is the information that supports the diagnostic, and the post treatment status is complementary and therefore optional.

1.3.5.1 Staging

Information that supports the evidence of the diagnostic can be condensed into the staging. Cancer staging determines the extent to which a cancer has developed by spreading. It takes into account the size of the tumor, whether it has invaded adjacent organs, how many lymph nodes it has spread to, and whether it has appeared in more distant locations. The stage will be expressed using the "American Joint Committee on Cancer" TNM clinical stage of the patient.

In order to express the three components of the TNM Clinical category (Tumor category, node category and metastasis category), it is required a main act with the code "TNM Breast Cancer Staging 254326001" related specifically with other three acts that will express the information of one of each the components.

The tumor category is used to describe the size of the original tumor and whether it has invaded nearby tissue. It is expressed with the code "Tumor Stage 37150800" with a value within the range of the value set "Breast TNM Tumor Category 2.16.840.1.113883.11.20.11.13" (e.g. "T4 category 65565005"). This act will be complemented with cancer staging method value set, from the possibilities "American Joint Commission on Cancer, Cancer Staging Manual, 6th edition neoplasm staging system 444256004" or "American Joint Commission on Cancer, Cancer Staging Manual, 7th edition neoplasm staging system 443830009".

The node category describes nearby lymph nodes that are involved. It is expressed with the code "Node Category Finding 385382003" with a value within the range of the value set "Breast TNM Node Category 2.16.840.1.113883.11.20.11.14" (e.g. "N0 category 62455006"). This act will be complemented with cancer staging method value set, from the possibilities cited before.

The metastasis category describes spread of cancer from one part of the body to another. It is expressed with the code "M - Category 277208005" with a value within the range of the value set "Breast TNM Metastasis Category 2.16.840.1.113883.11.20.11.15" (e.g. "M1 category 55440008"). This act will be complemented with cancer staging method value set, from the possibilities cited before.

1.3.5.2 Post Treatment Disease Status

The post treatment disease status represents the status of the breast cancer (in this case) after the treatment of the patient.

It is expressed with a single act, with the code "Post procedural recovery status 405178006" and a value from the value set "Disease Status Post Treatment 2.16.840.1.113883.10.20.30.4.1 DYNAMIC" (e.g. "Complete therapeutic response 399056007").

2 The basics of the HL7 v3 message generation

2.1 Basic structure of a HL7 v3 message

A HL7 v3 message or HL7 CDA is a XML based markup standard that specifies the structure and semantics of clinical documents in order to be exchanged. The CDA specifies that there is a structure part that relies on coding systems, like SNOMED or LOINC, and a textual part that ensures human interpretation of the document contents. A CDA document contain a header, with a set of minial require and a body. A CDA document contain a header, with a specified minimal required information, and a body that contains the information about the subject of the document, usually a patient. This section will introduce the mineral required elements that must be present in every CDA document.

The header describes the document itself, and its possible relationships to other documents. And it is composed by the following minimal set of elements:

- **ClinicalDocument**: All documents begin with this root element that contains the namespace declarations
    - **typeID**: References the CDA Release 2 specification and must contain the values root = "2.16.840.1.113883.1.3" (OID for HL7 Registered models) and extension = "POCD_HD000040" (unique identifier for the CDA Release 2 Hierarchical Description)
    - **id**: Unique identifier of the clinical document
    - **code**: Specifies the particular kind of document using the LOINC terminology.
    - **effectiveTime**: document creation time following the HL7 adoption of the ISO8601
    - **confidentialityCode**: contextual component where it is express the privacy of the document, from not classified as sensitive like public information, to extremely sensitive which presents a very high risk if disclosed without authorization
    - **recordTarget**: Represents the person whose char the document belongs to, typically a patient. Contains a nested element that specifies the role as well as the element that contains the target person information
    - **author**: Represents the humans and/or machines that authored the document
    - **custodian**: Represents the organization that is in charge of maintaining the document

There are other CDA optional elements that can be used at the header like: **relatedDocument**, **setId**, **componentOf**, **legalAuthenticator**, **templateId** and **documentationOf**.

The body of the CDA document is composed of one or more sections that can be nested, and which are related through a component relationship. Each section contains a title and text with a narrative block (human readable), as well as entries which convey the machine-computable semantics of the section.

A CDA body can be represented with a **nonXMLBody**, or a **structuredBody** element. A **nonXMLBody** contains a text element, which can be any kind of **mediaType**, such as image/tiff, application/pdf, text/html, text/plain, etc.

The **structuredBody** element contains at least one **section** elements that contain clinical information of the document. One of each section contain a **code** that identifies it, and a title that will display the heading of the section. The section contains as well a **text** element that is a simple structural markup. The section may contain an **entry** element, which carry the computable semantics of the report. A key component of this mechanism is the **entryRelationship**, which create relationships between entries, nested or referenced. There are existing implementation guides that prescribe patterns of entries and assign template identifiers to support validation against requirements of this implementation guides. These implementation guides are out of the scope of this document due to the specificity of each implementation.