ICT-2010-270253

INTEGRATE

Driving excellence in Integrative Cancer Research through Innovative Biomedical Infrastructures

STREP
Contract Nr: 270253

Deliverable: 3.2 Initial Proposal for the Core Dataset

Due date of deliverable: (31-06-2012)
Actual submission date: (22-08-2012)

Start date of Project: 01 February 2011 Duration: 36 months

Responsible WP: Philips

Revision: submitted

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Service	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (excluding the Commission Services)	

0 DOCUMENT INFO

0.1 Author

Author	Company	E-mail
Anca Bucur	Philips	anca.bucur@philips.com
Jasper van Leeuwen	Philips	jasper.van.leeuwen@philips.com
David Perez del Rey	UPM	dperez@infomed.dia.fi.upm.es
Raul Alonso Calvo	UPM	ralonso@infomed.dia.fi.upm.es
Philippe Henebert	IJB	philippe.hennebert@bordet.be
Jérôme Lhaut	IJB	jerome.lhaut@bordet.be
Marianne Paesmans	IJB	Marianne.paesmans@bordet.be
Alexandre Irrthum	BIG	alexandre.irrthum@bordet.be
Marion Maetens	BIG	marion.maetens@bordet.be
Lina Pugliano	BIG	lina.pugliano@bordet.be

0.2 Documents history

Document version #	Date	Change
V0.1	01.05.2012	Starting version, template
V0.2	01.05.2012	Definition of ToC
V0.3		First complete draft
V0.4		Integrated version (send to WP members)
V0.5		Updated version (send PCP)
V0.6	15.06.2012	Updated version (send to project internal reviewers)
Sign off	01.08.2012	Signed off version (for approval to PMT members)
V1.0	20.08.2012	Approved Version to be submitted to EU

0.3 Document data

Keywords	Core dataset, standard terminologies, ontologies
Editor Address data	Name: Anca Bucur Partner: Philips Address: High Tech Campus 34, 5656AE Eindhoven, The Netherlands Phone: +31 40 27 49709 Fax: E-mail: anca.bucur@philips.com
Delivery date	22.08.2012

0.4 Distribution list

Date	Issue	E-mailer
		Gisele.Roesems@ec.europa.eu
		fp7-integrate@listas.fi.upm.es

Table of Contents

1 Introduction

The INTEGRATE project aims to support a novel research approach in oncology through the development of innovative biomedical infrastructures enabling multidisciplinary collaboration, management and large-scale sharing of multi-level data, and the development of new methodologies and of predictive multi-scale models in cancer. The INTEGRATE infrastructure will bring together heterogeneous multi-scale biomedical data generated through standard and novel technologies within post-genomic clinical trials and seamlessly link to existing research and clinical infrastructures, such as clinical trial systems, eCRFs, and hospital EHRs, in order to enable a range of innovative applications.

The project also aims to make relevant steps towards semantic interoperability. To be able to reuse previous efforts in data sharing, modeling and knowledge generation, and to access relevant external sources of data and knowledge it is beneficial to adhere whenever possible to widely-accepted standards and ontologies. The use of standards will also support wide scale adoption of our solutions.

An important objective of this project is to build tools that facilitate efficient the execution of post-genomic multi-centric clinical trials in breast cancer. A range of such tools aim to support recruitment through the automatic evaluation of the eligibility of patients for trials based on matching the characteristics of the patient population required by the trial to the patient data available for instance in the hospital EHR.

Clinical trials are key instruments in clinical research that enable the validation of research hypotheses turning them into evidence that can be applied in wide clinical care. The population suitable to be enrolled in a trial is described by a set of free-text eligibility criteria that are both syntactically and semantically complex and whose automatic evaluation in order to assess the eligibility of a patient for a set of trials is a challenging task.

As criteria describe characteristics of the eligible patient population that need to be matched against the data items that are known for an individual patient, this task would be facilitated by the ability to identify the semantic entities that sufficiently describe the meaning of the criteria and by establishing links to relevant available data, stored for instance in an EHR system. Building these links (mappings) is a partially manual process and it is beneficial to be able to reuse them whenever possible across trials and systems.

This report focuses first on the analysis of the semantics of the eligibility criteria of clinical trials based on widely used medical ontologies. We identify the subsets of the ontologies that sufficiently capture the content of the eligibility criteria of trials in the clinical domain of interest which is breast cancer and compare with trials in cancers other than breast and in the cardiovascular domain. Our approach to identifying relevant subsets of ontologies relies on the annotation with ontology concepts of large corpuses of clinical trial eligibility criteria. We prioritize relevant concepts based on their frequency in the breast cancer subset and on their co-occurrence in trials in other domains.

Next, we identify sets of concepts that are relevant in other relevant data sources in the research domain (Clinical Report Forms, CT databases) and in the care domain (Electronic Health Records). In all these cases we map these concepts to selected ontologies that are relevant for the clinical research domain. An analysis of the selected ontologies is also provided.

Finally we also analyze the syntactic patterns that occur in eligibility criteria of trials and work towards a formalization of these patterns allowing automatic evaluation. By analyzing a large set of breast cancer clinical trials we derived a set of patterns that capture typical structure of conditions, pertaining to syntax and semantics. We qualitatively analyzed their expressivity and evaluated coverage using regular expressions, running experiments on a few thousands of clinical trials also related to other diseases.

We evaluate whether our modular approach for the selection of the sets of concepts based on the clinical domain is scalable and feasible. Selecting subsets instead of using entire ontologies facilitates the linkage of the clinical trial criteria to the actual patient records. The definition of mappings or other processing steps for entire ontologies is not feasible because of the sizes of the ontologies.

1.1 Aim of this Document and Outline

At the centre of the semantic solution that links trial descriptions to the patient data is the core dataset as in [5]: Soundly defined and agreed-upon clinical structures consisting of standard-based concepts, their relationships, quantification etc., that together sufficiently describe the clinical domain. To maximize reuse we evaluate the ability of several ontologies to capture the semantics of the criteria. It is of interest to identify the subsets of the ontologies that cover the meaning of the criteria in relevant clinical domains, the sizes of these subsets, the frequencies of concepts across trials and the overlap between subsets that describe criteria of trials in different domains. This information enables us to evaluate the effort required for the implementation of mappings, the priorities in building these mappings and the scalability of our solution with the number of trials and the extensibility to other domains.

We aim to capture the semantics of the clinical terms by standard terminology systems such as SNOMED-CT[4], MedDRA[5] and LOINC[6], which are widely used in the clinical domain. The scalability of the solution needs to be achieved by modularization, e.g. instead of aiming at inclusion of the complete SNOMED terminology we will identify a core subset that covers the chosen clinical domain and the datasets in our repositories. In the process of identifying the core dataset and the corresponding mapping tools, we need to allow for easy extension of this core dataset when the inclusion of new concepts becomes necessary (e.g. when adding new trials).

The selection of this core dataset is both clinical domain- and application-specific. Our first application area is clinical trial recruitment. To support automatic assessment of the suitability of patients for trials we need to be able to capture the semantics of the eligibility criteria and to evaluate if those are satisfied by the available patient data. Therefore, to define the initial core dataset we start from data sources that are relevant in this context: on one side the eligibility criteria and the CRFs of clinical trials and on the other the Electronic Health Records.

After identifying the concepts that define the semantics of the criteria we need to bind those to the information model of the system containing the patient data. As the development of these mappings is a time consuming and partially manual process it is important to minimize the effort required. Therefore, we need to evaluate the sizes of the concept sets that are relevant and the ease of handling updates (e.g. adding new clinical trials, and incorporating changes/updates in the ontologies used or in the information models of the sources) and extensions to new clinical domains. These aspects are important to assess the feasibility of our solution. In this report we try to answer some of these questions by evaluating the semantic content of the trial eligibility criteria based on widely-used ontologies.

2 Identification of the Core Dataset Information in Clinical Trial Descriptions

An important objective of the INTEGRATE project is to build tools that support the efficient execution of post-genomic multi-centric clinical trials in breast cancer, which includes the automatic assessment of the eligibility of patients for available trials. The population suited to be enrolled in a trial is described by a set of free-text eligibility criteria that are both syntactically and semantically complex. At the same time, the assessment of the eligibility of a patient for a trial requires the (machine-processable) understanding of the semantics of the eligibility criteria in order to further evaluate if the patient data available for example in the hospital EHR satisfies these criteria.

This section presents an analysis of the semantics of the clinical trial eligibility criteria based on relevant medical ontologies in the clinical research domain: SNOMED-CT, LOINC, MedDRA. We detect subsets of these widely-adopted ontologies that characterize the semantics of the eligibility criteria of trials in various clinical domains and compare these sets. Next, we evaluate the occurrence frequency of the concepts in the concrete case of breast cancer (which is our first application domain) in order to provide meaningful priorities for the task of binding/mapping these ontology concepts to the actual patient data. We further assess the effort required to extend our approach to new domains in terms of additional semantic mappings that need to be developed.

2.1 Description of the Experiment and of the Dataset

In order to analyze the semantics of eligibility criteria of clinical trials we have selected a large set of trial descriptions out of those published on ClinicalTrials.gov, a service of the U.S. National Institute of Health. We have used ClinicalTrials.gov because this site is widely used by the clinical research community and the set of trials available is both comprehensive and representative for our applications.

We selected trials from three clinical domains: breast cancer, cancer other than breast cancer, and heart and blood diseases. TABLE I. indicates the number of trials in each of the three domains. The breast cancer corpus was selected as relevant because it is the first domain for which we will implement our semantic solution and trial recruitment tools. The second corpus, clinical trials that study cancer other than breast cancer, and the third, trials that investigate heart and blood diseases, will enable us to compare the semantics of the different domains and to evaluate our modular approach and the extensibility to a new clinical domain.

TABLE I. NUMBER OF TRIALS IN THE EVALUATION

Clinical domain		
Breast cancer	***Cancer other than breast***	***Heart and blood diseases***
4232	6691	12255

We extracted the eligibility criteria from these sets of trials and used a state of the art annotator to identify the ontology concepts present in these criteria. The annotator is available at BioPortal[1] and is developed by the National Center for Biomedical Ontology. The BioPortal annotation results include information such as the concept name, concept identifier and the UMLS[2] semantic type of the concept.

The annotator allows to select out of a library of approximately 300 biomedical ontologies those that are relevant for the user. We have selected SNOMED-CT, MedDRA and LOINC.

We extracted and analyzed the sets of ontology concepts that were found to link to items from the eligibility criteria of our selected collection of clinical trials and compared the result for the three

[1] http://bioportal.bioontology.org/

[2] http://www.nlm.nih.gov/research/umls/

clinical domains selected and the three medical ontologies. The results of this analysis are described in the next section.

2.2 Evaluation Results

The first step in the analysis was to identify the sizes of the sets of concepts that describe the semantics of a domain and how much of the entire ontology they represent. Our modular approach to semantic linkage would not work if a large part of those ontologies is relevant for the trial criteria, for instance because implementing semantic mappings for a large ontology such as SNOMED-CT (over 311 000 concepts in 2011) requires a huge effort and would not be feasible for our application.

Next, we have compared the subsets of concepts among the different domains and for the three ontologies to identify overlaps and extensions. This enables us to estimate the effort of implementing our solution for the initial domain of breast cancer and the ease of extending this solution to new domains.

Another aspect of interest is to compare trials in each domain and assess how similar the semantics of distinct trials are. A large degree of similarity (which would be expected) means that once implementing our solution for a sufficiently large set of trials, adding new trials requires little effort. It is also relevant to prioritize the concepts that are occurring most often.

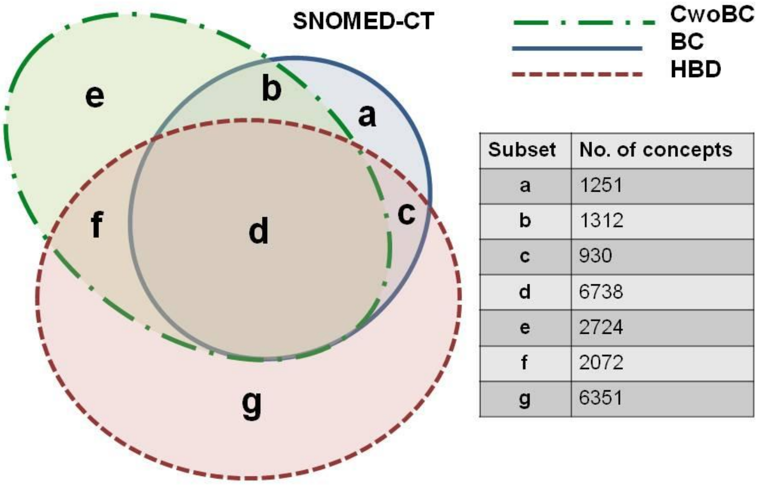
Finally, we investigated the most frequent semantic types that correspond to the concepts identified in the criteria. This additional information is relevant as it enables us to classify concepts with similar content or from similar sources.

2.2.1 Subsets of Concepts

The figures below compare the sets of relevant concepts for the three clinical domains and the three ontologies that we selected. In all cases the largest set is the one that is the overlap among the three domains, therefore concepts that are domain independent. The breast cancer corpus has a small subset that is specific for this disease (marked with “a” in the figure) and also relatively small subsets that constitute overlaps with each of the other domains. This makes our modular approach very feasible as a large amount of the concepts used in the semantic solution for breast cancer will be also relevant for other diseases.

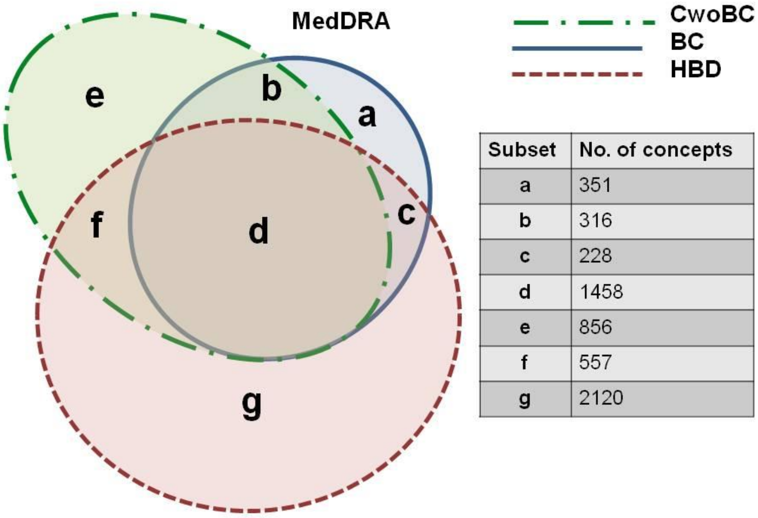
Extensions to new domains are also manageable as the additional sets of concepts are relatively small, even for completely different domains (e.g. extending from BC to HBD). This is especially the case for LOINC, where the module covering the concepts that are specific for HBD is half the size of the overlap with BC and CwoBC.

Also in absolute numbers, the sizes of the sets of concepts that capture the semantics of our domains are reasonable and support the implementation of our semantic solution and of the trial recruitment applications that will rely on it.



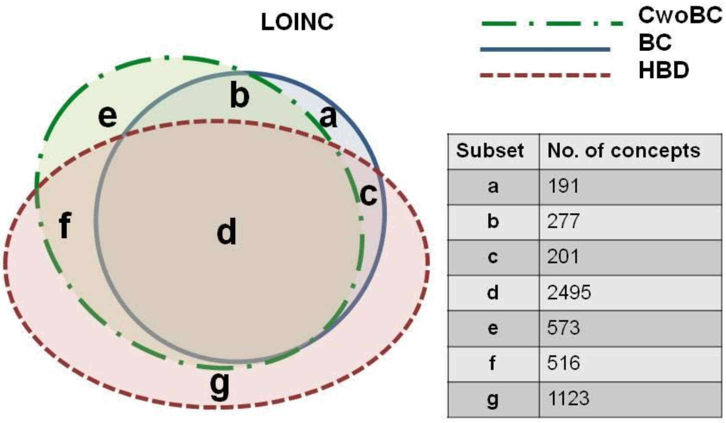
Subset	No. of concepts
a	1251
b	1312
c	930
d	6738
e	2724
f	2072
g	6351

Figure 1. Sets of SNOMED-CT concepts for breast cancer (BC), cancer other than breast cancer (CwoBC) and heart and blood disease (HBD)



Subset	No. of concepts
a	351
b	316
c	228
d	1458
e	856
f	557
g	2120

Figure 2. Sets of MedDRA concepts for breast cancer (BC), cancer other than breast cancer (CwoBC) and heart and blood disease (HBD



Subset	No. of concepts
a	191
b	277
c	201
d	2495
e	573
f	516
g	1123

Figure 3. Sets of LOINC concepts for breast cancer (BC), cancer other than breast cancer (CwoBC) and heart and blood disease (HBD)

TABLE II. indicates the ratio of the three selected ontologies that cover the semantics of our domains. It shows that a small percentage of the ontologies are sufficient to capture the content of the trial descriptions in the specific domains. For instance, in the case of breast cancer trials 3.2 % of SNOMED CT, 3.3% of MedDRA and 5.4% of LOINC were used in the annotation of our datasets.

TABLE II. RATIO OF THE ONTOLOGIES THAT CAPTURE THE SEMANTICS OF THE THREE SETS OF CLINICAL TRIALS

	SNOMED-CT	***MedDRA***	***LOINC***
BC	0.032	0.033	0.054
CwoBC	0.041	0.045	0.066
HBD	0.051	0.062	0.074

2.2.2 Reoccurrence of concepts across trials

In this section we investigate the semantic similarity among trials. Intuitively we expected that trials will have a large ratio of criteria that are similar, but that new trials do introduce new concepts. This is confirmed by Figure 4. that depicts for the three corpuses of trials and the three ontologies the distribution of concepts across trials. Only the top most frequent concepts are depicted and each concept is counted once per trial. In all cases there is a relatively small group of concepts that occur in a large number of trials and there is another group of concepts that are rare or unique for specific trials. To further illustrate this, TABLE III. provides for each selected clinical domain and ontology the average number of trials in which a concept occurs, the average number of trials for the top 100 most frequent concepts, and the average number of trials for the top 500 most frequent concepts.

To have an additional reference, we have also counted the number of ontology concepts that occur per trial. In the case of breast cancer we have concluded that there are on average 199 SNOMED-CT concepts per trial, 27 MedDRA concepts and 108 LOINC concepts.

These facts enable us to prioritize the implementation of semantic mappings starting with the concepts that occur often and demonstrate that the effort of adding new trials is low: Updates will be required, but the additional concepts that need to be mapped to relevant data are few. TABLE IV. and TABLE V. include examples of very frequent concepts of SNOMED-CT and MedDRA that occur in the BC dataset.

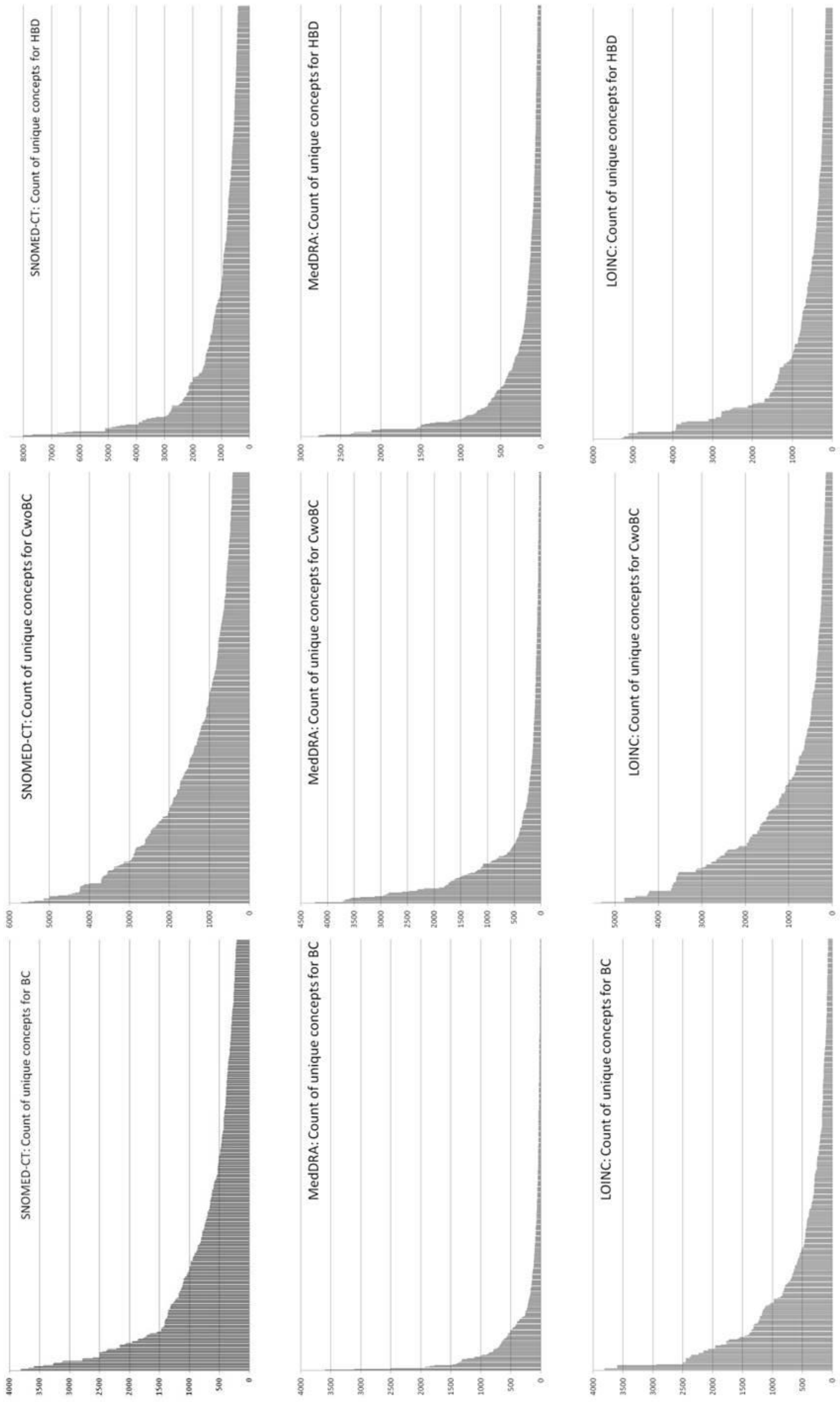


Figure 4. The reoccurrence of concepts across trials. The number of BC, CwoBC and respectively HBD trials (y-axis) that include the top 500 most frequently occuring concepts (x-axis) out of SNOMED-CT, MedDRA and LOINC. Concepts were counted once per trial.

TABLE III. STATISTICS OF THE REOCCURENCE OF ONTOLOGY CONCEPTS IN BREAST CANCER TRIALS

Reoccurrence of ontology concepts for BC (4232 trials)			
Statistics	***SNOMED-CT***	***MedDRA***	***LOINC***
Average number of trials for top 100 concepts	1835.57	632.61	1597.45
Average number of trials for top 500 concepts	756.63	163.01	510.11
Average number of trials for all unique concepts	51.865	36.98	89.31
Number of distinct concepts	10231	2353	3164
Concepts occuring in a single trial	3159	815	671

TABLE IV. FREQUENT SNOMED-CT CONCEPTS IN BREAST CANCER TRIALS (ALL OCCURENCES COUNTED)

Concept name	*Concept code*	*Number of occurrences*
Disease	64572001	13261
Neoplasm, malignant (primary)	86049000	11241
Entire breast	181131000	10254
Breast structure	76752008	10246
Therapeutic procedure	103733002	9958
Therapy	276239002	9956
Malignant neoplastic disease	363346000	9764
History of	392521001	8170
Study	224699009	7823
Malignant tumor of breast	254837009	5544
Antineoplastic chemotherapy regimen	69960004	5544
Drug therapy	182831000	5336

TABLE V. FREQUENT MEDDRA CONCEPTS IN BREAST CANCER TRIALS (ALL OCCURENCES COUNTED)

Concept name	*Concept code*	*Number of occurrences*
Cancer	10007050	9737
Breast cancer	10006187	6133
Chemotherapy	10061758	5320
Metastatic	10027474	3689
Carcinoma	10007284	2759
Surgery	10042609	2739
Radiotherapy	10037794	2612
Pregnant	10036586	2148
Metastases	10027476	2109
Creatinine	10011358	1956

2.2.3 Semantic Types

In this section we evaluate the UMLS semantic types of the concepts in our datasets as they can provide additional information about the semantics of the criteria and identify concepts that are similar. We compare the frequency in the sets of concepts of several semantic types that are relevant for our application domain. We have annotated with semantic types all the concepts identified in all three ontologies.

The tables below depict the most frequent semantic types for the three corpuses: BC, CwoBC and HBD. We can observe that there are little differences among the clinical domains in the hierarchy of semantic types to which most of the concepts belong.

TABLE VI. RATIO OF THE MOST FREQUENT UMLS SEMANTIC TYPES FOR THE BC DATASET (RELATIVE TO TOTAL NUMBER OF CONCEPTS)

Semantic type	***Ratio***
Neoplastic Process	0.055
Qualitative Concept	0.041
Therapeutic or Preventive Procedure	0.039
Body Part, Organ, or Organ Component	0.031
Laboratory Procedure	0.023
Finding	0.018
Pharmacologic Substance	0.012
Diagnostic Procedure	0.010
Quantitative Concept	0.006
Spatial Concept	0.005

TABLE VII. RATIO OF THE MOST FREQUENT UMLS SEMANTIC TYPES FOR THE CWOBC DATASET (RELATIVE TO TOTAL NUMBER OF CONCEPTS)

Semantic type	***Ratio***
Qualitative Concept	0.080
Therapeutic or Preventive Procedure	0.076
Neoplastic Process	0.065
Laboratory Procedure	0.049
Body Part, Organ, or Organ Component	0.041
Finding	0.032
Diagnostic Procedure	0.022
Pharmacologic Substance	0.022
Quantitative Concept	0.012
Spatial Concept	0.010

TABLE VIII. RATIO OF THE MOST FREQUENT UMLS SEMANTIC TYPES FOR THE HBD DATASET (RELATIVE TO TOTAL NUMBER OF CONCEPTS)

Semantic type	***Ratio***
Therapeutic or Preventive Procedure	0.044
Qualitative Concept	0.041
Body Part, Organ, or Organ Component	0.038
Finding	0.036
Laboratory Procedure	0.029
Pharmacologic Substance	0.020
Diagnostic Procedure	0.018
Quantitative Concept	0.015
Spatial Concept	0.011
Sign or Symptom	0.010

3 Identification of Core Dataset Information by Domain Experts

Other important sources for the identification of relevant concepts that should be included in the core dataset are the clinical trial databases and the trial CRFs. As a starting point, to have an overview of the INTEGRATE core dataset concepts relevant for these sources, domain experts of the consortium (oncologists, trial managers, bioinformaticians, data managers, etc.) manually identified concepts that are relevant for their trials and mapped them on the three ontologies that were selected: SNOMED-CT, MedDRA and LOINC. These ontologies were considered the best choices due to their wide adoption in the clinical research domain.

3.1 Overview

In this section we analyze the concepts that were manually identified and evaluate the correspondent terms in the relevant ontologies. For the terms that were found, we evaluate whether in the case of SNOMED-CT post-coordination could be used. We concluded that most terms are found in the ontologies or can be generated by post coordination, but there are still cases of terms that are relevant and cannot be found in the ontologies. This situation was to be expected as domains evolve and our clinical users work at the cutting edge of innovation in clinical research.

In cases where there is a lack of pre- or post-coordinated terms, extensions to the core dataset are required. This should be handled in such a way as to maintain the adherence of our solution to the standards and the ontologies selected and to keep the cost and effort of updates as low as possible. At the same time, our results based on the analysis of the trials on ClinicalTrials.gov show that the numbers of new concepts that are introduced by new trials are relatively small and updates can be handled with little effort in the presence of the right processes.

3.2 Evaluation of the Concepts

In TABLE IX. we present the concepts identified by the users and for each of them available corresponding concepts in SNOMED-CT, MedDRA and LOINC, when those could be found.

TABLE IX. CONCEPTS MANUALLY-IDENTIFIED BY EXPERTS AND THE CORRESPONDING ONTOLOGY CONCEPTS

Name	Snomed-CT Concept (# Code)	MedDra Concept (# Code)	LOINC Concept (# Code)
Concepts related to Patient Data			
Date of Birth	Date of birth (Code 184099003)	🟥	Birth date (code 21112-8)
Gender	Gender (Code 263495000)	Gender related factors, 10018057	Gender patient (code 21840-4)
ECOG grade	ECOG performance status (423740007)	🟥	ECOG performance status grade (42800-3)
Operable	Operable (Code 76234009)	🟥	🟥
Histologic type	Histologic type (Code 371441004)	Histology, 10062005	Histologic type (44638-5) breast cancer

Grade	Grade (Code 103421006)	🟥	Grade pathology value, 59542-1 Grade pathology sytem, 59541-3
Tumour size	Tumor size (Code 263605001)	🟥	Tumor size.collaborative staging, 42079-4
Positive Lymph Nodes	Positive (Code 10828004)	🟥	Regional lymph nodes positive, 21893-3
Metastasis	Secondary malignant neoplastic disease (128462008)	Metastasis, 10062194	M stage of distant metastasis, 44666-6
Metastasis location	Secondary malignant neoplastic disease (128462008) Associated topography (Code 116677004)	🟥	Distant metastasis site, 44667-4
HER2 status ICH	Human epidermal growth factor receptor 2 gene detection by immunohistochemistry (Code 433114000)	🟥	HER2, 48676-1
HER2 status FISH ratio	Human epidermal growth factor receptor 2 gene detection by fluorescence in situ hybridization (Code 434363004) Ratio (Code 118586006)	🟥	HER2/CEP17, 49683-6
HER2 status FISH copy number	Human epidermal growth factor receptor 2 gene detection by fluorescence in situ hybridization (Code 434363004)	🟥	HER2, 48675-3
Ki67	Rapidly proliferating cell marker (Code 259981004)	🟥	🟥
Estrogen Receptor	Estrogen receptor (Code 23307004)	Estrogen receptor assay, 10054060	Estrogen receptor, 14130-9
Progesterone receptor	Progesterone receptor (Code 61078009)	Progesterone receptor assay, 10054056	Progesterone receptor, 10861-3
Neutrophils (ANC) [2]	Absolute (Code 56136002) Neutrophil count (Code 30630007)	Neutrophils, 10029380	Neutrophils by Automated count, 751-8 Neutrophils by Manual count, 753-4

Platelets	Platelet count (61928009)	Platelet count, 10035525	Platelets, 777-3
Hemoblobin	Measurement of total haemoglobin concentration (441689006)	Hemoglobin, 10019481	Hemoglobin, 717-9
Bilirubin	Serum bilirubin measurement (166610007)	Bilirubin, 10004683	Bilirubin, 1972-9
ALT/SGPT	SGPT - blood level (250636007)	Alanine aminotransferase, 10001546	Alanine aminotransferase, various
AST/SGOT	SGOT measurement (45896001)	Aspartate aminotransferase, 10003476	Aspartate aminotransferase, various
ALP	Alkaline phosphatase measurement (88810008)	Alkaline phosphatase, 10001674	Alkaline phosphatase, various
Creatinine	Creatinine level (365756002)	Creatinine, 10011358	Creatinine, various
LVEF	Left ventricular ejection fraction (Code 250908004)	Left ventricular ejection fraction, 10069170	Ejection fraction, various
Pregnant	Patient currently pregnant (Code 77386006)	Pregnant, 10036586	Are you currently pregnant, 66174-4 Have you ever been pregnant, 63892-4
Contraception	Contraception (Code 13197004)	Contraception, 10010808	Contraception risk, 42836-7
Menopausal status	? (premenopausal, postmenopausal or menopausal statusses are easy to find)	Menopausal, 10027296	Are you currently using any over-the-counter - herbal, natural, or soy-based - preparations for hormone replacement or to treat post-menopausal symptoms, 64649-7
Previous non breast cancer	Can be postcoordinated using BEFORE, CANCER, BREAST and a negation that should be included in the model. For negation, see http://sage.wherever.org/cresources/cresources.html#HL7 or use EXCEPT FOR (5185003)	🟥	🟥
Inflammato	Inflammatory carcinoma	Inflammatory breast	🟥

ry breast cancer	of breast (254840009)	cancer, 10021974	🟥
Tumour Laterality	Neoplasm (Code 108369006) Laterality (Code 272741003)	🟥	Site & laterality & morphology override flag, 22003-8
Multifocal / Multicentric	Multifocal (Code 524008) Multicentric (Code 255206009)	Extrasystoles multifocal, 10015859	🟥
Serious Cardiac disorder	Serious (Code 42745003) Heart disease (Code 56265001)	Cardiac disorder, 10061024	Cardiac output alteration, 28149-3
Serious Medical Condition	Serious (Code 42745003) Disease (Code 64572001)	Medical observation, 10053047	🟥
Current infection	Current (Code 15240007) Infectious disease (Code 40733004)	Infection, 10021789	🟥
Serious Mental disorder	Serious (Code 42745003) Mental disorder (Code 74732009)	Mental disorder, 10061284	🟥
Serious Gastrointestinal disorder	Serious (Code 42745003) Disorder of digestive tract (Code 84410009)	Disorder gastrointestinal, 10013225	Gastrointestinal alteration, 28111-3
Enrolled in other trial	🟥	🟥	🟥
Ongoing medications	Current or specified time (Code 410512000) Drug or medicament (Code 410942007)	🟥 🟥	🟥 🟥
Ethnicity	Ethnic group finding (Code 397731000)	🟥	Ethnicity, 54120-1

Concepts related to Clinical Trials or Trial Arms

Trial type	Trial == Not found Type (Code 410656007)	🟥	Trial name, 42796-3 Trial design, 35513-1
Randomized	Random (Code 255226008)	🟥	🟥
Sponsor	🟥	🟥	🟥
Clinical endpoint	Clinical (Code 58147004) Endpoint == Not found	🟥	Clinical information, 55752-0

Region of world	Geographical and/or political region of the world (Code 223496003)	🟥	Birthplace
Arm type	🟥	🟥	🟥
Drug	Drug or medicament (Code 410942007)	Drug-drug pharmacodynamic interaction, 10065993	Drugs identified, various
Tissue availability	Tissue specimen (Code 119376003) Availability of (Code 103328004)	🟥	Tissue type, 55073-1

Concepts related to Samples			
Collection	Specimen collection (Code 17636008)	🟥	🟥
Type	Type (Code 410656007)	🟥	Type, various
Parent sample	Part of (Code 123005000) or some other relational concept	🟥	🟥
Location	Location within hospital premises (Code 224884006)	🟥	Location of Care Area Assessment (CAA) information, 58196-7
Quality	Quality (Code 263496004)	🟥	Quality control and quality assurance section, 35522-2
Quantity	Quantity (Code 246205007)	🟥	🟥
Conseted analyses	Consented (Code 441898007) Analysis (Code 272389005)	🟥	🟥

Concepts related to Drugs			
Molecule	Molecule (Code 290005005)	🟥	🟥
Class	Class (Code 277046005)	🟥	Class, various
Target	🟥	🟥	🟥

3.2.1 Concepts Identified in the Relevant Ontologies

In this section we summarize the results of the analysis and specify the numbers of concepts that were found in each of the three ontologies, those that could be generated in SNOMED-CT through post-coordination and those that were not found. This shows that only few concepts could not be found in the selected ontologies and that these ontologies offer a good coverage for our domain of interest.

TABLE X. NUMBER OF CONCEPTS MANUALLY-IDENTIFIED BY EXPERTS USING THE THREE SELECTED ONTOLOGIES

Vocabulary	Found	Not Found	Post-cordination
SNOMED	40	7	12
MedDra	25	34	0
LOINC	38	21	0

TABLE XI. PERCENTAGE OF CONCEPTS FOUND, NOT FOUND AND FOUND BY POST-COORDINATION ON SNOMED

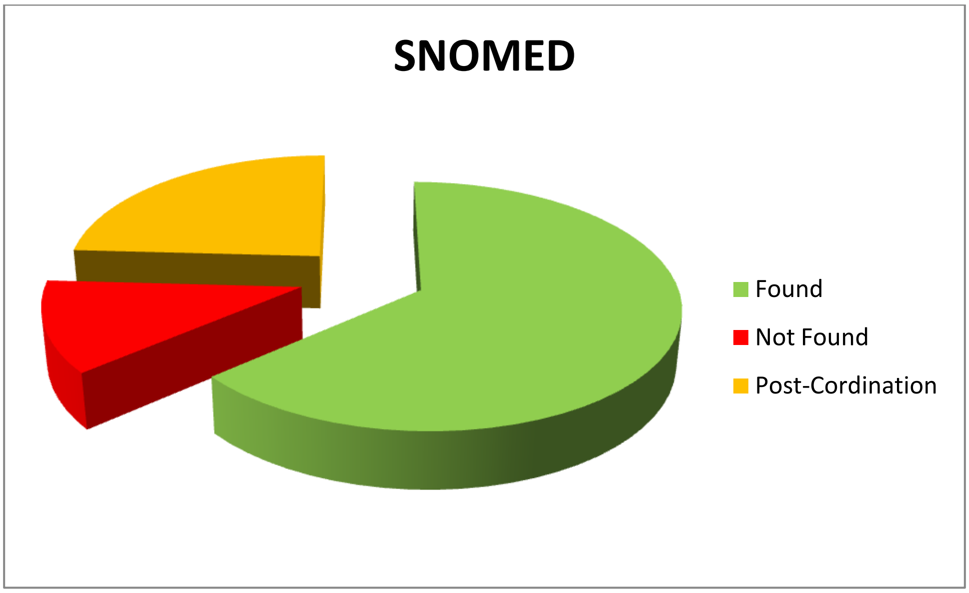


TABLE XII. PERCENTAGE OF CONCEPTS FOUND AND NOT FOUND ON MEDDRA

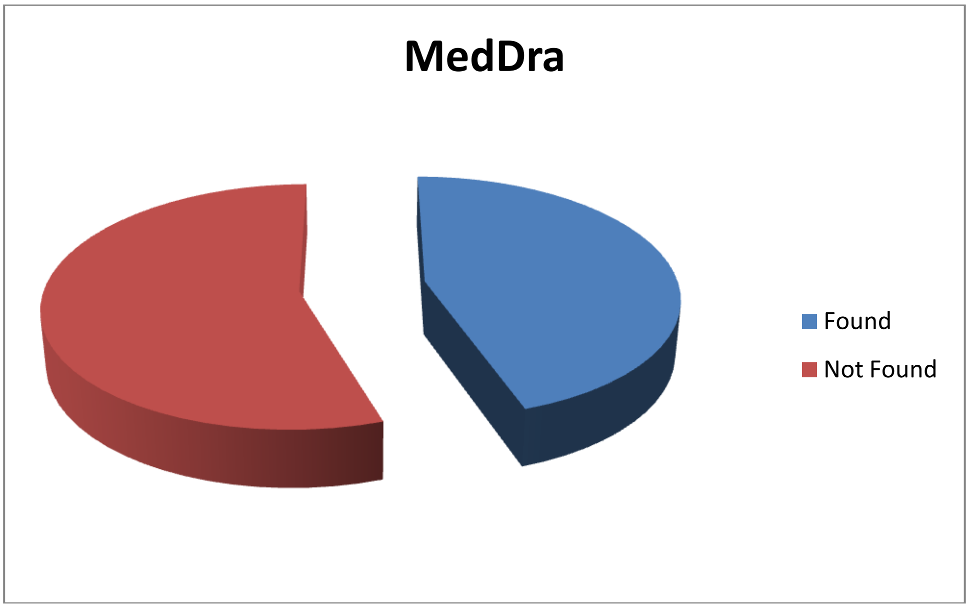
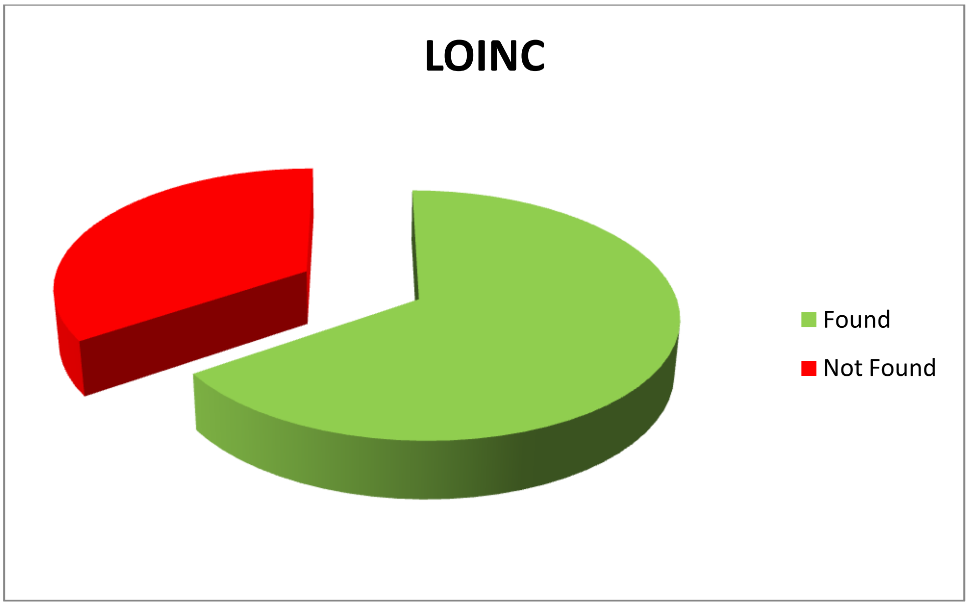


TABLE XIII. PERCENTAGE OF CONCEPTS FOUND AND NOT FOUND IN LOINC



4 Identification of Core Dataset Information in Electronic Health Record Data

At the core of the INTEGRATE semantic interoperability layer lies a standards-based semantic core dataset which is linked to the canonical information models that represent the other relevant systems. The semantic core dataset together with the devised mappings will enable the linkage between the data sets and models stored and managed by the INTEGRATE infrastructure and the data in the relevant external systems. This enables the users, when desired, to securely link predictive models of response to therapies back to the actual patients participating in the trials, providing this way a quick transfer of results to the treating clinician and improving patient outcome.

As an important goal of INTEGRATE is to semantically link the data in our environment to the data in the Electronic Health Record System, we need access to such data for the development of the semantic layer of INTEGRATE.

4.1 Overview

Three sources of patient data from TOP trial were used to identify concepts for the initial proposal of the INTEGRATE core dataset, extracted from the EHR system:

- Pathology data (Diamic)
- Pharmaceutical data (Infohos)
- Laboratory data (Glims)

A total of 2464 HL7 acts were required to store data from 50 patients in the INTEGRATE common data model. The identification of core dataset concepts was performed using SNOMED as a baseline, but including LOINC concepts for laboratory tests as extensions. In fact, there was a higher number of different LOINC concepts, although the majority of acts were coded using an SNOMED code. A limited number of SNOMED codes were sufficient to represent the required semantics.

4.2 Evaluation of the Concepts

The following tables presents the different concepts from SNOMED and LOINC, identified within the EHR data from TOP trial.

TABLE XIV. PRE-COORDINATED SNOMED CONCEPTS PRESENT AT EHR DATA FROM TOP TRIAL

SNOMED Code	Concept Name
326830005	Aclarubicin (product)
4590003	Adenocarcinoma, metastatic (morphologic abnormality)
34608000	Alanine aminotransferase
104485008	Albumin
88810008	Alkaline phosphatase

45896001	Aspartate aminotransferase
42351005	Basophils/100 leukocytes
55235003	C reactive protein
71878006	Calcium
104589004	Chloride
42525009	Coagulation surface induced
103220009	Coagulation tissue factor induced actual/Normal
396451008	Coagulation tissue factor induced.INR
113075003	Creatinine
35300007	Daunorubicin (product)
71960002	Eosinophils/100 leukocytes
349848009	Epirubicin (product)
54706004	Erythrocyte mean corpuscular hemoglobin
37254006	Erythrocyte mean corpuscular hemoglobin concentration
104133003	Erythrocyte mean corpuscular volume
14089001	Erythrocytes
69480007	Gamma glutamyl transferase
80274001	Glomerular filtration rate/1.73 sq M.predicted
28317006	Hematocrit
441689006	Hemoglobin
108786002	Idarubicin (product)
82711006	Infiltrating duct carcinoma
767002	Leukocytes
19225000	Lorazepam (product)
74765001	Lymphocytes/100 leukocytes
38151008	Magnesium
108791001	Mitoxantrone (product)
67776007	Monocytes/100 leukocytes
73572009	Morphine (product)
72495009	Mucinous adenocarcinoma (morphologic abnormality)
30630007	Neutrophils/100 leukocytes
30566004	Noninfiltrating intraductal papillary adenocarcinoma (morphologic abnormality)
104867005	Phosphate
75672003	Platelet mean volume
61928009	Platelets
59573005	Potassium
74040009	Protein
104934005	Sodium
387713003	Surgical procedure (procedure)
108507005	Tramadol (product)
263605001	Tumor size

105011006	Urea nitrogen
116079002	Valrubicin (product)
96231005	Zolpidem (product)

TABLE XV. POST-COORDINATED SNOMED CONCEPTS PRESENT AT EHR DATA FROM TOP TRIAL

SNOMED Code	Concept Name
76752008	Breast structure (body structure)
78615007	with laterality
7771000	left
76752008	Breast structure (body structure)
78615007	with laterality
24028007	right

TABLE XVI. LOINC CONCEPTS PRESENT AT EHR DATA FROM TOP TRIAL

LOINC Code	Concept Name
2862-1	Albumin
13980-8	Albumin/Protein.total
2865-4	Alpha 1 globulin
13978-2	Alpha 1 globulin/Protein.total
2868-8	Alpha 2 globulin
13981-6	Alpha 2 globulin/Protein.total
1798-8	Amylase
2871-2	Beta globulin
13982-4	Beta globulin/Protein.total
1959-6	Bicarbonate
1975-2	Bilirubin
1968-7	Bilirubin.glucuronidated+Bilirubin.albumin bound
1989-3	Calcidiol
6875-9	Cancer Ag 15-3
2039-6	Carcinoembryonic Ag
2093-3	Cholesterol
2085-9	Cholesterol.in HDL
55440-2	Cholesterol.in LDL
19080-1	Choriogonadotropin
2106-3	Choriogonadotropin (pregnancy test)
3243-3	Coagulation thrombin induced
2132-9	Cobalamins
2157-6	Creatine kinase
788-0	Erythrocyte distribution width

4537-7	Erythrocyte sedimentation rate
2243-4	Estradiol
2276-4	Ferritin
3255-7	Fibrinogen
2283-0	Folate
2284-8	Folate
15067-2	Follitropin
2874-6	Gamma globulin
13983-2	Gamma globulin/Protein.total
2345-7	Glucose
4542-7	Haptoglobin
13950-1	Hepatitis A virus Ab.IgM
13954-3	Hepatitis B virus little e Ag
5196-1	Hepatitis B virus surface Ag
13955-0	Hepatitis C virus Ab
48345-3	HIV 1+O+2 Ab
728-6	Hypochromia
2498-4	Iron
2502-3	Iron saturation
2532-0	Lactate dehydrogenase
10501-5	Lutropin
735-1	Lymphocytes.variant/100 leukocytes
715-3	Normoblasts
51579-1	Normoblasts/100 cells
2692-2	Osmolality
2839-9	Progesterone
2842-3	Prolactin
14196-0	Reticulocytes
4679-7	Reticulocytes/100 erythrocytes
11579-0	Thyrotropin
3021-3	Thyroxine binding globulin
3024-7	Thyroxine.free
3034-6	Transferrin
3040-3	Triacylglycerol lipase
2571-8	Triglyceride
3053-6	Triiodothyronine
3051-0	Triiodothyronine.free
3055-1	Triiodothyronine/Triiodothyronine uptake index
3084-1	Urate

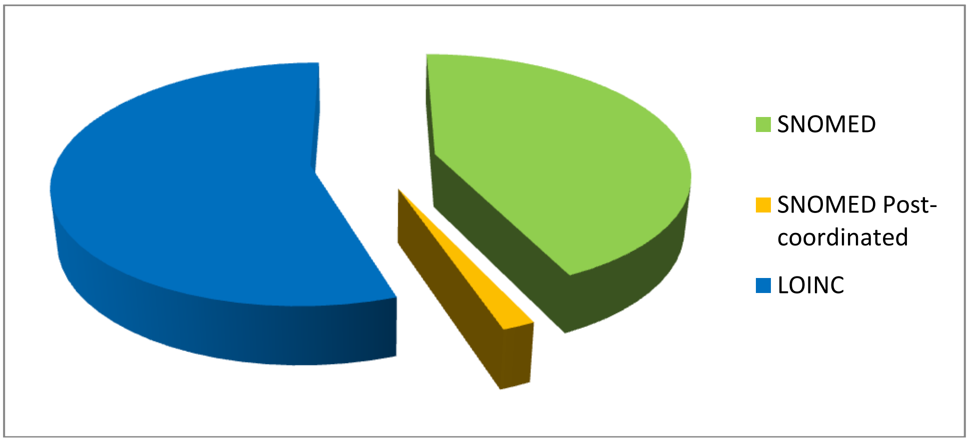
4.2.1 Concepts Identified in the Relevant Ontologies

In this section we summarize the results of the analysis and specify the numbers of concepts that were found in SNOMED and LOINC. At the moment there are no extensions required to represent all the data.

TABLE XVII. TOTAL NUMBER OF CONCEPTS FROM SNOMED AND LOINC PRESENT AT EHR DATA FROM TOP TRIAL

SNOMED	SNOMED Post-coordinated	LOINC	Total
49	2	63	114

TABLE XVIII. PERCENTAGE OF CONCEPTS FROM SNOMED AND LOINC PRESENT AT EHR DATA FROM TOP TRIAL

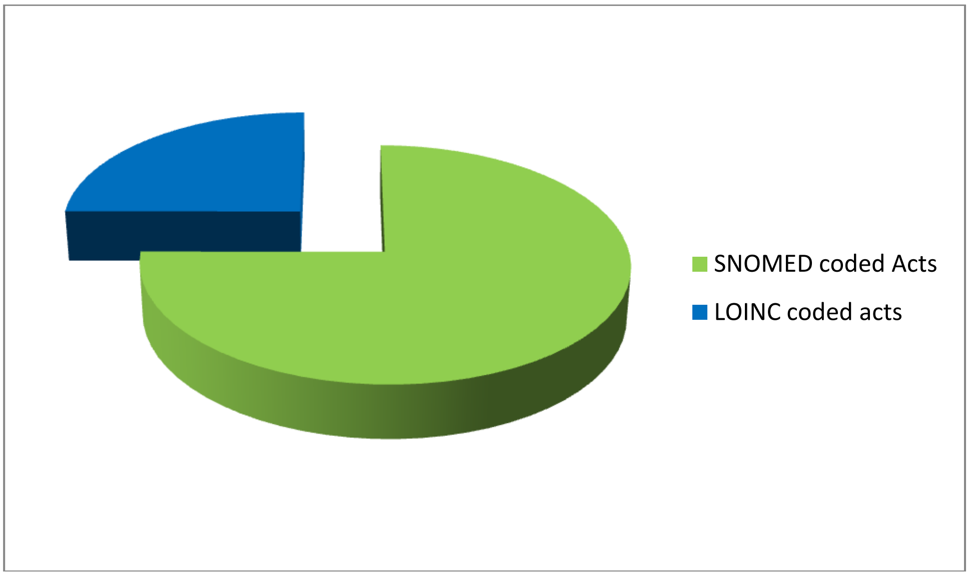


Although we needed a high number of different LOINC codes for lab tests, the majority of the data was encoded using a limited set of SNOMED concepts.

TABLE XIX. TOTAL NUMBER OF ACTS CODED WITH SNOMED AND LOINC CONCEPTS TO STORE EHR DATA FROM TOP TRIAL

SNOMED coded Acts	LOINC coded acts	Total
1849	615	2464

TABLE XX. PERCENTAGE OF OF ACTS CODED WITH SNOMED AND LOINC CONCEPTS TO STORE EHR DATA FROM TOP TRIAL



5 Identification of Context Patterns in Clinical Trial Criteria

In this section we address the issue of detecting the context patterns in eligibility criteria. In [4], by analyzing a large set of breast cancer clinical trials we derived a set of patterns that capture typical structure of conditions, pertaining to syntax and semantics. We qualitatively analyzed their expressivity and evaluated coverage using regular expressions, running experiments on a few thousands of clinical trials also related to other diseases. A detailed evaluation of the patterns in terms of coverage and of the pattern detection algorithm we described in [7]. We concluded that the patterns selected cover the language of eligibility criteria that describe the context of the core dataset concepts (which were the focus of the rest of this report) to a large extent and may serve as a semi-formal representation.

5.1 Introduction

With the objective to facilitate the automatic assessment of trial eligibility, we propose a formalization of the criteria that enables the extraction of the machine-processable semantics of the criteria based on which to evaluate the match to the corresponding patient data.

The formalization method (semi-)automatically interprets the semantics by identifying in each criterion two relevant types of entities: concepts that express the core meaning and modifiers (syntactic patterns) that provide the context of the criterion. This method supports automated matching and reasoning for applications such as determining patient eligibility for clinical trials or designing eligibility criteria for new trials to improve study feasibility. In [4] we have identified relevant syntactic patterns and evaluated their coverage for a large set of eligibility criteria and their expressivity. Additionally, we define a multi-dimensional classification of criteria that aims to support the information extraction of required patient data, scoping, and semantic search in the context of applications such as trial matching, protocol design and feasibility.

5.2 The Pattern Detection Method

Our approach to formalizing the eligibility criteria involves several steps, depicted in Figure 5. To develop the method we first build a knowledge base by processing a large number of eligibility criteria of existing trials to extract syntactic and semantic structures that appear (with different frequencies) in criteria and are relevant for the selection of the required patient data and for the evaluation of patient's eligibility. We start with initial pre-processing of free text of eligibility criteria, then we identify patterns in eligibility criteria that provide the context of the criterion.

The knowledge base of patterns was developed by processing a large number of eligibility criteria of existing trials. We used existing NLP frameworks and extracted syntactic and semantic structures that appear (with different frequencies) in criteria and are relevant for the selection of the required patient data and for the evaluation of patient's eligibility. The patterns are described in detail in following section.

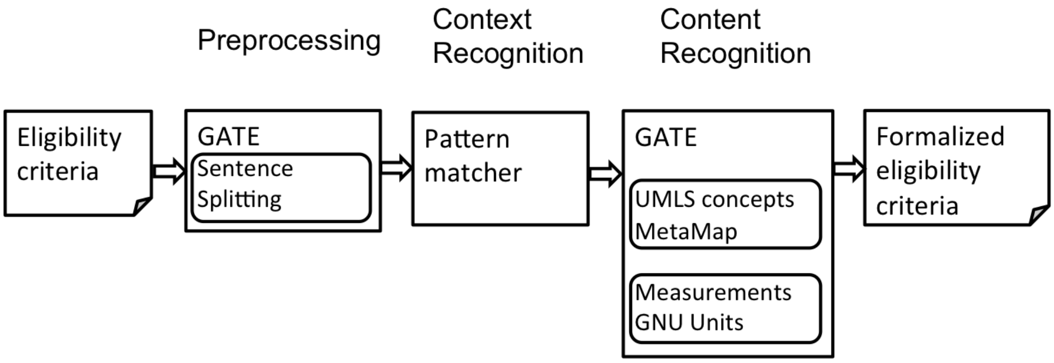


Figure 5. Pipeline of formalization steps for clincal trial eligiblity criteria

The following example illustrates the approach. The pattern detection algorithm would recognize in the eligibility criterion "Has received chemotherapy within the past 14 days" the pattern "prior () within ()". Further, its content is annotated and as a result "chemotherapy" is recognized as UMLS concept with identifier C0013216, 14 days as measurement with value='14' and unit='day'.

5.3 Definition of Representative Patters

By inspecting the definition of eligibility criteria of clinical trials, we have observed that the language of these criteria is regular and there is a significant reuse across trials. It inspired us to define a set of patterns and analyze to which extent they capture the language used to define eligibility criteria. We started an informal development process by extracting eligibility criteria from the description of all available breast cancer trials (3905). Further we focused the analysis on a randomly selected subset, containing approximately few hundreds of trials. To identify common ways of expression we manually grouped conditions by similar subject (demographic information, disease characteristic, prior- concurrent treatment) or similar syntax. We noticed that criteria differ in the level of complexity. Some are formulated as atomic phrases e.g. 'Not pregnant', others as complex sentences e.g. 'Brain metastases allowed provided they have been treated with surgery.' We aimed to define patterns covering both groups, incrementally extending a set of patterns.

The patterns represent the canonical form, and can be instantiated in numerous ways, e.g. a pattern "No history of ()" correspond to both criteria: "Must not have a history of CNS metastasis" and "No prior metastatic malignancy".

We have also observed that criteria consist of semantic concepts relevant in the clinical domain, modifiers describing the context of the criterion and common English words. The method of defining patterns, developed during the formalization process, inspired by observed concrete examples of eligibility criteria, can be summarized as follows. In order to cover sentence structure we started from basic forms e.g. "must be receiving ()" and added corresponding negated versions "can not be receiving", as well as past tense, both positive and negative e.g. "must have received ()" and future if applicable. Secondly we extended the resulting basic forms with common specifications, which restrict for example time frame, purpose of a treatment or co-occurrences. If applicable, these were combined. An example of a pattern containing two specifications: time frame and exclusions is 'more than () since prior () except for ()', capturing criteria like 'More than 6 months since prior endocrine therapy, except tamoxifen'.

Additionally, we defined patterns that capture atomic phrases, covering value restrictions for chosen parameters, expressed by arithmetic comparison or enumerated values, and their negations. Patterns that capture atomic phrases can be nested in the patterns reflecting context in the sentence structure. As a result an initial set of 135 different patterns were defined, which after two rounds of evaluation were extended to 165.

TABLE XXI. EXAMPLES OF PATTERNS FOUND IN STANDARD ELIGIBILITY CRITERIA

Dimension	**Example of pattern**	**Example of condition**
Time frame	At least () since prior()	At least 3 weeks since prior steroids
Exclusions	No prior () except for ()	No prior malignancy, except for adequately treated basal cell.
Value restrictions	T () stage; Age above ()	T2; Age >18
Confirmation	confirmed by ()	No metastasis to brain (confirmed by CT or MRI)
Medical content		
Menopausal status	Post- menopausal	Postmenopausal women
Pathology data	margins must be clear	Resected margins histologically free of tumor.
Molecular data	Known gene mutation	Documented BRCA1/2 mutation.

5.4 Classification Dimensions

Based on experiments, we identified that the following properties would improve the automatic reasoning capabilities with trial eligibility criteria:

a) Criteria specific:
b) **Medical content:** cancer type, treatment, pathology, clinical, molecular, imaging, laboratory, informed consent, etc.
c) **Data source of medical content:** patient history, family history, findings, conclusions, discharge diagnosis, current medications, laboratory, imaging, and trial specific data.
d) **Time independent status:** present, absent, conditional (*'Multifocal breast tumors allowed if all foci are ER-negative'*), not selective (*Any N stage)*.
e) **Temporal status:** historical, current, planned.
f) **Variability and controllability** as proposed in[2]: stable, variable, controllable, subjective.

To support trial feasibility and matching we propose some additional dimensions:

f) Trial specific:
g) **Importance:** user defined hierarchy reflecting the order in matching and the possibility of relaxing the original criterion.
h) **Adjustability:** reflects the possibility of relaxing the original criteria.
i) Institution specific:
f) **Selectiveness:** estimates the ratio of patients from the population that satisfy the criterion.
g) **Decidability:** indicates the likelihood of finding the required information in the clinical information system (for the automatic evaluation of eligibility criteria) when the criterion is satisfied.

TABLE XXII. EXAMPLES OF PATTERNS AND THEIR CLASSIFICATION

Eligibility criteria	**Medical Content**	**Time independent status**	**Temporal status**	**Variability and controllability**
Histologically confirmed invasive breast cancer	Cancer type	Present	Current	Variable
Known hormone receptor status	Pathology data	Present	Current	Controllable
No malignant neoplasms within 10 years, unless curatively treated	Clinical data	Conditional	Historical	Variable
Negative pregnancy test, within 2-weeks prior to randomization	Clinical data	Absent	Current	Controllable
Platelet count ≥ 100 x 10^9/L	Laboratory data	Present	Current	Variable
No prior treatment for primary invasive breast cancer	Therapy	Absent	Historical	Stable (if yes) / variable
No metastasis (M0) (isolated supraclavicular node involvement allowed)	Pathology data	Conditional	Current	Variable
Must be receiving trastuzumab.	Treatment	Present	Current	Variable
Known carrier of BRCA1 or BRCA2 mutation	Molecular	Present	Current	Stable

Eligibility criteria	Adjustability	Importance	Selectiveness	Decidability

Histologically confirmed invasive breast cancer	No	Essential	High	High
No metastasis	By exception e.g.: No metastasis (isolated supraclavicular node involvement allowed)	High	Medium	High
Platelet count ≥ 100 x 10^9/L	By value e.g.: Platelet count ≥ 90 x 10^9/L	Medium	Medium	High
Age < 60	By value	Medium	Low	High
No history of significant psychiatric disorders	By specification	Medium	High	Low

5.5 Evaluation of the Patterns

Evaluation of the formalization approach with respect to the detection of syntactic patterns in eligibility criteria is the main topic of [7]. The evaluation addresses several aspects: The precision and recall of the pattern detection algorithm and the assessment of the coverage of our set of syntactic patterns for the selected domain. The evaluation was performed manually using a subset of patterns and randomly selected 66 trials from ClinicalTrials.gov.

The algorithm for pattern detection is based on regular expressions. In total we defined 468 regular expressions corresponding to the 165 patterns. The algorithm processes eligibility criteria delimited using GATE[3] sentence splitter. Each sentence can correspond to more than one pattern.

From the set of patterns identified in the sentence, the algorithm chooses only those that cover the longest phrases, and ignores patterns capturing segments subsumed by others. For example in the sentence 'No other concurrent hormonal therapy, including steroids', it identifies two patterns 'no concurrent ()' and 'no concurrent () including ()', from which it selects only the latter because it more closely reflects the content and meaning of the criterion. In addition, it recursively searches for nested patterns. In the sentence: 'No history of other malignant neoplasms except for curatively treated nonmelanoma skin cancer or surgically cured carcinoma of the cervix in situ' the algorithm first identifies the pattern 'no prior () except for ()' and, second, the one nested in the second parameter 'recovery from ()'.

The results of applying the algorithm are the subject of our evaluation as described in the paper. In this section we summarize the results. Due to the significant manual effort required for the evaluation, we decided to focus on a selected subset of clinical trials and patterns, described next.

[3] http://gate.ac.uk/

5.5.1 Subset of Clinical Trials

We tested our method with clinical trials criteria from the large public repository ClinicalTrials.gov, using trials that specify breast cancer as a study condition. Our main focus in INTEGRATE lies here because this is the domain of our clinical partners, whose expertise will be crucial in further steps of the research. From the available clinical trials we randomly selected 1%.

5.5.2 Subset of patterns

We have selected 20 patterns out of the 165 for the evaluation: the 10 most frequent and the 10 most complex. The selection of most frequent patterns was based on the number of their occurrences in the eligibility criteria in the total corpus of over 3 thousand breast cancer clinical trials. The selection of the most complex patterns was based on the number of pattern variables (i.e. pattern "no ()" has 1 variable, pattern "no () within () except for ()" 3) and their availability in the selected subset of clinical trials. We distinguished the most complex patterns to verify whether the performance of the pattern detection algorithm depends on the complexity of the patterns.

5.5.3 Evaluation of the pattern detection algorithm

The set of patterns contains 165 patterns, which reflect the typical constraints put on the patient data. The patterns were defined in the iterative process of assessing and improving the expressivity of entire set.

We evaluated the pattern identification algorithm in terms of precision and recall and analyzed the results of the annotation of sentences of the selected set of eligibility criteria. We manually verified whether the patterns detected by the algorithm were indeed the best match from our set, and whether the algorithm has found all of them.

5.5.4 Summary of the results

The average precision for the group of most complex patterns is significantly higher than for the group of most frequent ones (0.98 vs. 0.83), while recall is lower (0.86 vs. 0.99). This finding confirms our intuition that the algorithm performs better in the correct identification of complex phrases. It should be noticed that the most frequent patterns account for almost 40% of all defined patterns, therefore the focus should be placed on preventing errors related to them, unless we develop an application focused on particular kinds of eligibility criteria.

The score of the annotation indicates the average extent to which a pattern chosen by the algorithm covers the details of the best matching pattern corresponding to the criteria. Within detailed inspection of the results, we observe that the detected elements of criteria indicate that in some cases even using suboptimal patterns can lead to correct filtering of patients. However, the opposite also can happen. The majority of mistakes are caused by failing to recognize the broader context, strengthening conditions, time independent status and then weakening conditions. An example of misinterpreted context is the recognition of a pattern: "History of ()" in a sentence "Prior systemic therapy in the adjuvant setting is not considered a regimen." which has only an explanatory role. The focus needs to be on preventing errors connected to misinterpreting the context, which would deteriorate both precision and recall of finding eligible candidates.

6 Analysis of the core dataset candidate ontologies

6.1 Introduction

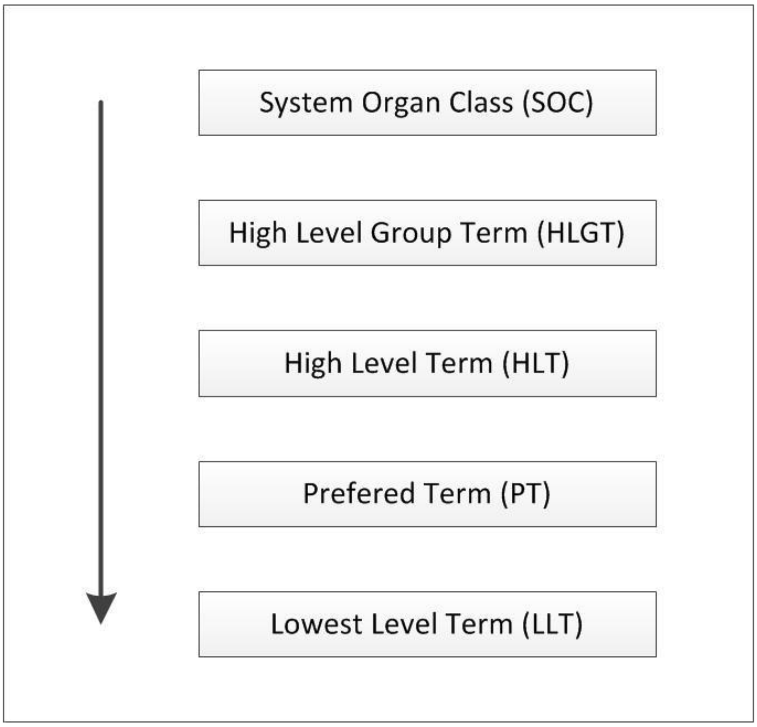
Three taxonomies, widely adopted in the area, have been selected to build the initial proposal of the INTEGRATE Core Dataset: SNOMED, MedDRA and LOINC (mentioned at D2.1). The main properties for each terminology have been analyzed regarding: structure, location, release format, and access. We also provide a final comparison table in order to facilitate the observation of the core datasets candidate characteristics.

6.2 Medical Dictionary for Regulatory Activities (MedDRA)

Developed by the IFPMA (International Federation of Pharmaceutical Manufacturers and Associations), MedDRA is a medical terminology developed in order to facilitate the sharing of the information about medical products used by humans. MedDRA terms refer to diseases, diagnoses and reactions and results to classify information related with adverse events associated to the use of biopharma and other medical products on humans. Nevertheless, and due to its open philosophy, its use is growing worldwide into many new areas as clinical research, beginning to be a standard for a lot of scientific regulatory authorities.

The structure of this vocabulary is hierarchical, i,e, terms "owned" by a sequence of predecessors terms, noting that one term could be preceded by more than one father-term. The hierarchy is composed by six levels:

TABLE XXIII. HIERARCHY OF MEDDRA



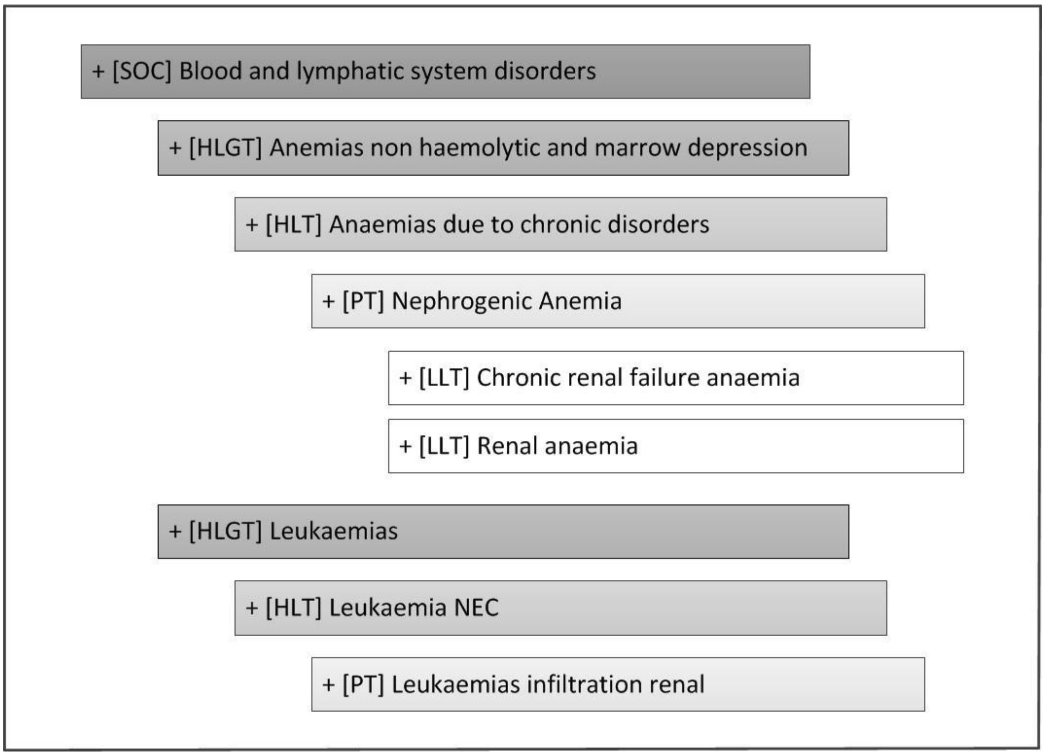
SOC represents the broadest concept; PT a single unique medical concept, and LLT a synonym or a lexical variant of a PT. Each term content at least one word, and are tagged with a code number that starts with 10000001 alphabetically, i.e, a term example could be “Urticaria” which is tagged with the code 100046735.

Regarding the origin of each of the terms that take part of the taxonomy, MedDRA is composed of terms from:

- COSTART (5th edition)
- WHO-ART (98:3)
- J-ARTS (1996)
- HARTS (Release 2.2)
- ICD-9
- ICD-9-CM (4th Revision)

A purchase license is required if it is intended to be used within a commercial software; true not-for-profit organizations qualify for a Basic subscription, for example an educational institution or a direct patient care provider, as an hospital planning to use MedDRA as a reference tool.

TABLE XXIV. EXAMPLE OF MEDDRA HIERARCHY



+ [SOC] Blood and lymphatic system disorders

+ [HLGT] Anemias non haemolytic and marrow depression

+ [HLT] Anaemias due to chronic disorders

+ [PT] Nephrogenic Anemia

+ [LLT] Chronic renal failure anaemia

+ [LLT] Renal anaemia

+ [HLGT] Leukaemias

+ [HLT] Leukaemia NEC

+ [PT] Leukaemias infiltration renal

MedDRA can be easily accessed by their users with the help of the Web-Based Browser at https://www.meddrabrowser.org/dsnavigator/ and the MedDRA Desktop Browser at http://www.meddramsso.com/subscriber_download_tools_browser.asp which facilitate searching for terms on the hierarchy.

6.3 Systematized Nomenclature of Medical – Clinical Terms (SNOMED-CT)

Property of the International Health Terminology Standards Development Organization (IHTSDO) since 2007; SNOMED-CT was born by the combination and expansion of the taxonomies SNOMED-RT, developed by the Colllege of American Pathologies (CAP), and CTV3, created by the National Health Service (NHS) of the United Kingdom.

Nowadays, SNOMED-CT is considered to be the most important clinical terminology, thanks to its precision and highly comprehension data. In addition, this taxonomy allows its users to tag, index and store clinical information; facilitating the correct management of medical media. Its usability has been an important help to professionals working with EHRs; becoming adopted as the standard clinical terminology for many institutions.

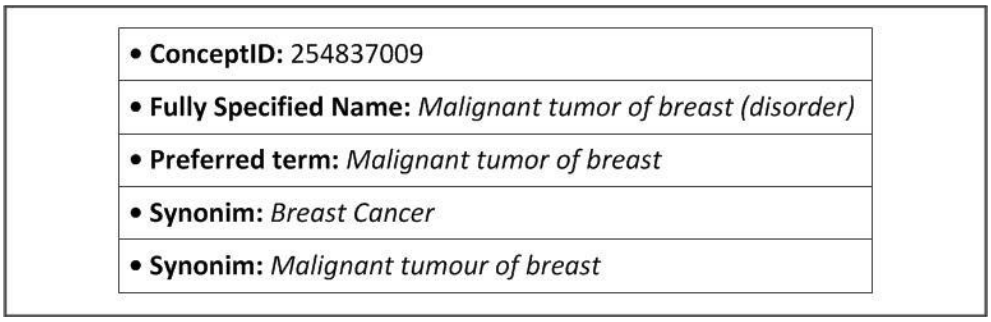
SNOMED-CT is composed of around a million of clinical meaning concepts identified by a single and unique number. Each concept has associated a few descriptions that describe different properties. Those descriptions could be:

- Fully Specified Name: A unique way to name and denominate the concept.

- Preferred Term: The common phrase/term used by clinics to name the concept.

- Synonym: Additional phrases/terms that could represent the concept.

TABLE XXV. SNOMED-CT TERM PROPERTIES

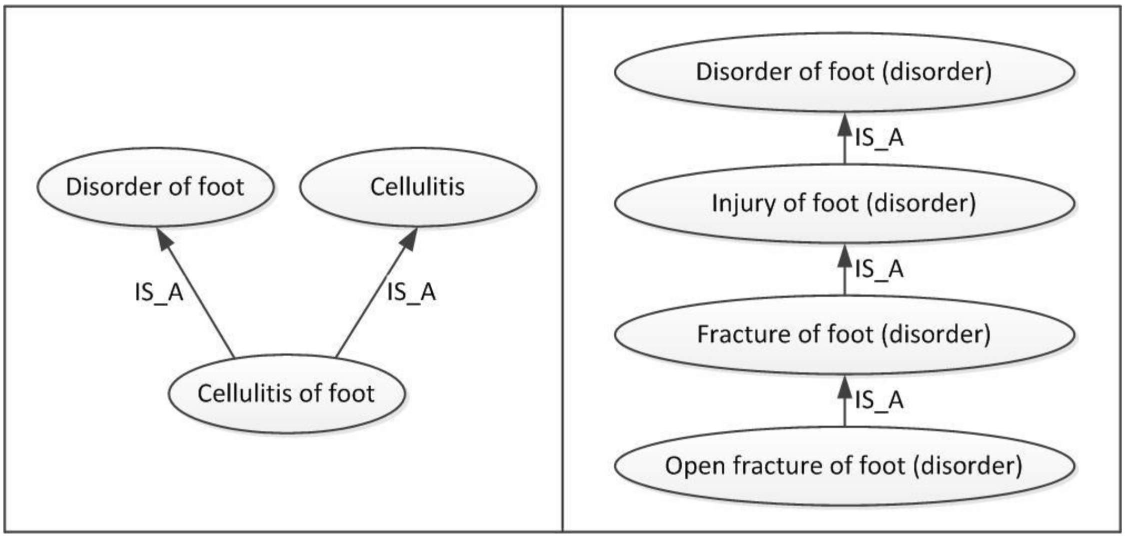


• **ConceptID:** 254837009
• **Fully Specified Name:** *Malignant tumor of breast (disorder)*
• **Preferred term:** *Malignant tumor of breast*
• **Synonim:** *Breast Cancer*
• **Synonim:** *Malignant tumour of breast*

Each concept in SNOMED-CT is logically defined through is relationships to other concepts. In fact, this vocabulary have two possible types of relationships:

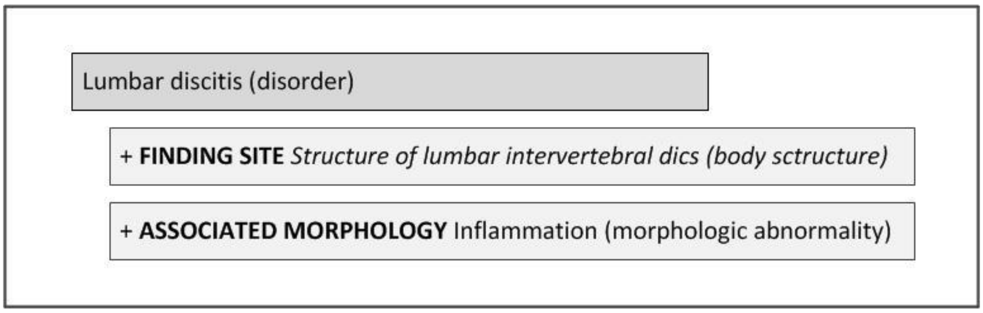
- IS-A relationships: Every concept has at least one IS_A relation to a super type concept and can have more than one IS_A relationship to other concepts. In that last case, the concept will have parent concepts in more than one sub-hierarchy.

TABLE XXVI. EXAMPLE OF SNOMED-CT RELATIONSHIPS


Disorder of foot
Cellulitis
IS_A
IS_A
Cellulitis of foot
Disorder of foot (disorder)
IS_A
Injury of foot (disorder)
IS_A
Fracture of foot (disorder)
IS_A
Open fracture of foot (disorder)

- Attribute relationships: Allow the logical representation of the meaning of a concept. They define the semantics of the elements and help to differentiate them from other similar concept definitions, including their own super-types and sub-types.

TABLE XXVII. EXAMPLE OF ATTRIBUTE RELATIONSHIPS



Lumbar discitis (disorder)

+ **FINDING SITE** *Structure of lumbar intervertebral dics (body sctructure)*

+ **ASSOCIATED MORPHOLOGY** Inflammation (morphologic abnormality)

Regarding availability of SNOMED-CT, the taxonomy is open for research purposes, although some restrictions apply to commercial products and depending on the country. There are also a set of free and useful web browsers on the internet, being specially interesting the SNOMED-CT core browser developed by the Virginia-Maryland Regional College, http://snomed.vetmed.vt.edu/sct/menu.cfm, and the one facilitated by the NCI,http://nciterms.nci.nih.gov/ncitbrowser/pages/vocabulary.jsf?dictionary=SNOMED%20Clinical%20Terms.

6.4 Logical Observation Identifiers Names and Codes (LOINC)

LOINC was developed to provide a standard for identifying clinical information in electronic reports. The LOINC vocabulary provides a set of universal names and ID codes for identifying laboratory and clinical test results in the context of existing HL7, ASTM E1238, and CEN TC251 observation report messages.

The LOINC codes are mainly intended to identify test results and clinical observation. Other fields in the LOINC message can transmit, for example, the identity of the source laboratory or other special details about the sample. A formal, distinct and unique name (composed by six parts) is given to each LOINC component term. The unique name has the following syntax:

<1. Analyte/component>:<2. kind of property of observation or measurement>: <3. time aspect>:<4. system (sample)>:<5. scale>:<[optional] 6. method>

Where:

- The name of the component or analyte measured (e.g., glucose, propranolol)

- The property observed (e.g., substance concentration, mass, volume)

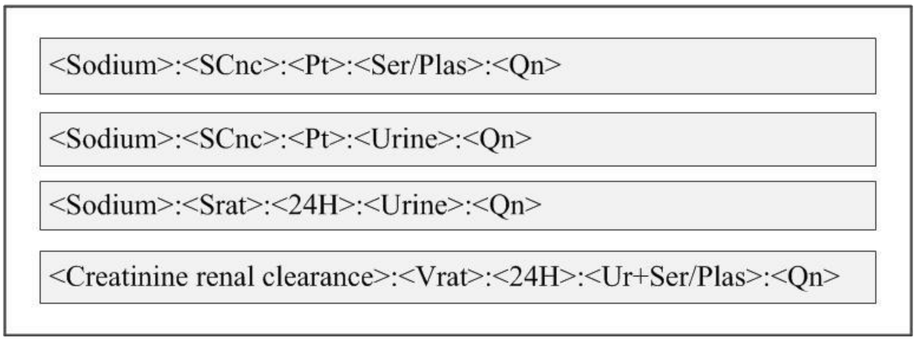
- The timing of the measurement (e.g., is it over time or momentary)

- The type of sample (e.g., urine, serum)

- The scale of measurement (e.g., qualitative vs. quantitative)

- The method of the measurement (e.g., radioimmunoassay, immune blot).

TABLE XXVIII. EXAMPLES OF LOINC NAMES



<Sodium>:<SCnc>:<Pt>:<Ser/Plas>:<Qn>

<Sodium>:<SCnc>:<Pt>:<Urine>:<Qn>

<Sodium>:<Srat>:<24H>:<Urine>:<Qn>

<Creatinine renal clearance>:<Vrat>:<24H>:<Ur+Ser/Plas>:<Qn>

LOINC is available as a Microsoft Access database file or as a tab-delimited text file. In order to search over LOINC, the Regenstrief Institute provides a Windows-based mapping utility called the RELMA (http://loinc.org/relma) to facilitate searches through the LOINC database and to assist efforts to map local codes to LOINC codes. The RELMA package includes the LOINC table and is available as a Windows-based mapping utility. A web search application is available at http://search.loinc.org/.

TABLE XXIX. COMPARISON OF THE THREE CORE DATASET CANDIDATES FOR INTEGRATE

	MedDRA	**SNOMED**	**LOINC**
Objective	- Facilitate the sharing of the information about medical products used by humans. - Classify information related with adverse events associated to the use of biopharma and other medical products on humans.	- To contribute to the improvement of patient care through the development of systems to accurately record health care encounters. - To deliver decision support to health providers.	- Facilitate the exchange and pooling of results for clinical care and research. - Identify laboratory and clinical test results.
Content	Diseases, diagnoses	All clinical areas:	Laboratory &

	and reactions and results.	Diseases, findings, microorganisms, pharmaceuticals.	Clinical content. For example, chemistry, hematologic, microbiology or clinical observation info.
Sources	COSTART (5th edition) WHO-ART (98:3) J-ARTS (1996) HARTS (Release 2.2) ICD-9 ICD-9-CM (4th Revision)	ICD-9-CM ICD-03 ICD-10 LOINC OPCS-4	Non-defined
Structure	Hierarchical: 6 levels	IS-A hierarchy with a non-defined number of levels	None
Tools	- Web Browser - Desktop Browser	- NVCI Term Browser - VTSL Core Browser	- RELMA - LOINC Browser
License	A purchase license is required if it is intended to be used within a commercial software	Free for research – License depending on the country	Free

7 Conclusions

In this report we have described our work of identifying the relevant concepts that enable us to scope the core dataset for our semantic solution. In the first version the identification of the core dataset considered three sources: Automatic evaluation out of eligibility criteria of clinical trials, manual identification carried out by the domain experts (clinicians, trial managers, bioinformaticians, molecular biologists, etc.) based on their experience but also using as source the trial data in the available systems, and automatic identification based on the relevant patient data in the EHR system.

First we have focused on the evaluation of the semantics of the eligibility criteria of clinical trials. In the context of developing applications supporting efficient execution of clinical trials it is essential to assess whether our modular semantic linkage approach is applicable to this domain. This requires to decide whether the semantics of the eligibility criteria can be captured by widely-used medical ontologies and to estimate the effort required to semantically link the eligibility criteria to the relevant patient data (for example preserved in an EHR) to enable the decision of whether the patient satisfies the criteria.

The ontologies selected were SNOMED-CT, MedDRA and LOINC which are widely used in the clinical domain and when used by our solutions could support scalability and adoption. We have identified the relevant subsets of these ontologies that capture the semantics of the eligibility criteria of clinical trials in selected clinical domains.

Another important question we have answered is of extendibility. Our main focus in the INTEGRATE project is breast cancer, but we aim to design solutions that can be extended and applied to other clinical domains. Therefore we evaluated and compared the sets of concepts that capture the semantics of different clinical domains: breast cancer, cancer other than breast cancer, and heart and blood disease.

The analysis of the concepts that are specific to a domain or occur across various domains let us modularize the sets of concepts that are relevant for a particular group of trials. We identified the subset of concepts that exclusively occur in eligibility criteria related to one of the three domains, those that are shared among trials in various clinical domains.

We have relied on the annotation of a large collection of clinical trials using the NCBO's BioPortal annotator. Our findings indicate that relatively small subsets (in terms of number of concepts) of the ontologies are required to capture the semantics of the eligibility criteria. It was also shown that the semantic overlap among clinical domains is very large for all ontologies considered, therefore once developed for a particular domain a large part of the mappings can be reused when extending the solution to a new domain. The additional sets of concepts that are specific to those domains are relatively small and the implementation of the new mappings is feasible.

The frequency of the concepts, their reoccurrence across various trials and their uniqueness for particular types of trials informs the selection of the concept sets that cover the meaning of the criteria. These statistics guide the process of linking the concepts to the data items in the patient records by building the necessary mappings.

We have concluded that the reuse of concepts across trials is very significant, with a relatively small number of concepts that occur in many trials. Therefore, they can be prioritized in the implementation of mappings. We can capture a large part of the semantics of the trials with a relatively small number of concepts that sufficiently describe the content of the eligibility criteria.

The infrequent concepts that are specific to single trials are also manageable and it is most efficient for the implementation of the semantic solution to only add those when a trial containing them is entered into the system. The long tail of the graph indicates that the sets of concepts identified will not be complete and will grow with new trials, but the high overlap across trials makes the effort of handling updates for new trials low.

We have also evaluated the UMLS semantic types of the concepts as these can provide additional hints about the semantics of the criteria and can be used in the semantic solution to reason about the criteria at a higher level of abstraction. We compared the frequency in the sets of concepts of several semantic types that are relevant for our application domain.

8 Related work

In this report we focus mainly on clinical trials in breast cancer and present an analysis of the semantics of the eligibility criteria of trials based on widely-adopted medical ontologies: SNOMED-CT[4], MedDRA[5] and LOINC[6].

SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms) is a clinical vocabulary focused on accurately recording health care encounters and the associated electronic health information exchange. Although SNOMED-CT is sometimes criticized, it has a significant uptake in clinical practice, such as its use in HL7[7] messaging. MedDRA focuses on the regulatory process of drug development and is a medical vocabulary that is used by regulatory bodies and the regulated biopharmaceutical industry for data entry, retrieval, evaluation and display. MedDRA is used in clinical trials for reporting adverse events. LOINC (Logical Observation Identifiers Names and Codes) has the purpose to facilitate the exchange and pooling of results for clinical care, outcomes management, and research. LOINC provides universal identifiers for laboratory and other clinical observations and it is a preferred code set for HL7 for laboratory test names in transactions between health care facilities, laboratories, laboratory testing devices, and public health authorities.

With respect to the selection of domain-specific parts of ontologies, in [1] the subset of UMLS that is relevant to describe breast cancer treatment was identified in order to facilitate the development of clinical decision support systems. While the general idea is comparable, the purpose and the method were different. As background knowledge the concepts from medical guidelines were used, considered at the decision points of selecting a suitable treatment for a patient. The guideline concepts were manually mapped to the SNOMED-CT concepts and the subset obtained was automatically expanded via the ontology hierarchy and the UMLS semantic network.

A significant body of research has focused on the general problem of formalization of eligibility criteria and on recruitment for clinical trials, including (semi)-automatic trial matching. In [2] an extensive overview of existing solutions and approaches is provided. In previous work [3] we have analyzed the eligibility criteria of clinical trials and we have identified relevant syntactic patterns that occur in trial criteria. These patterns are modifiers that provide the context of the criterion and while they do not express the semantics of the criterion and cannot be linked to actual patient data for evaluation of eligibility, they provide the context of the criterion. We have evaluated the coverage of these patterns for a large set of eligibility criteria and their expressivity.

In [4] an analysis has been carried out to estimate the coverage provided by SNOMED-CT for clinical research concepts that represented by the items present on case report forms (CRFs). The authors also evaluated the semantic nature of those concepts relevant to post-coordination methods. The dataset included a total of 17 CRFs developed by rheumatologists conducting several longitudinal, observational studies in the clinical domain of vasculitis. From the CRFs a total set of 616 (unique) items were identified. Each unique data item was classified as either a clinical finding or procedure. The items were coded by the presence and nature of SNOMED CT coverage and manually classified into semantic types by 2 coders.

[4] http://www.ihtsdo.org/SNOMED CT/
[5] http://www.meddramsso.com
[6] http://loinc.org/
[7] http://www.hl7.org/

In [5] we introduce a scalable, modular and pragmatic approach to achieving semantic interoperability. We believe that interoperability in healthcare can be achieved gradually on specific domains and by making use whenever possible of existing standards. This is also the approach that we take in the INTEGRATE project for a well-defined clinical domain which is clinical trials in breast cancer. As presented in this paper, we identify those modules of ontologies that are relevant in this domain and in our semantic solution we will implement mappings for those specific concepts. This facilitates efficient further extensions to other domains of relevance and easy reuse of tools. A gradual approach to interoperability is well supported in literature. In [6] it is stated that “regardless of the type of vision one may develop, semantic interoperability is not a phenomenon to be expected over night”. The group of experts conclude that semantic interoperability in healthcare requires a large number of changes at both the technical and the use case level, and that even in that vision no full semantic interoperability or a complete harmonization of either EHR models or terminologies can be expected.

9 Bibliography

[1] R. Vdovjak, B. Claerhout, and A. Bucur, "Bridging the Gap between Clinical Research and Care - Approaches to Semantic Interoperability, Security & Privacy", in HEALTHINF 2012, pp. 281-286.

[2] K. Milian, Z. Aleksovski, R. Vdovjak, A. ten Teije, and F. van Harmelen, "Identifying disease-centric subdomains in very large medical ontologies", in Proceedings of the AIME'09 workshop on Knowledge Representation for Healthcare (KR4HC09). 2009; pp. 41-50.

[3] C. Weng, S.W. Tu, I. Sim, and R. Richesson, "Formal representation of Eligibility Criteria: A Literature Review", in Journal of Biomedical Informatics, 2010.

[4] K. Milian, A. ten Teije, A. Bucur, and F. van Harmelen, "Patterns of clinical trial eligibility criteria", in Proceedings of the AIME'11 workshop on Knowledge Representation for Healthcare (KR4HC11), lecture notes AI, 2011.

[5] R.L. Richesson, J.E. Andrews, and J.P. Krischer, "Use of SNOMED CT to Represent Clinical Research Data: A Semantic Characterization of Data Items on Case Report Forms in Vasculitis Research", in JAMIA 13(5), 2006, pp. 536-546.

[6] V.N. Stroetmann et al., "Semantic Interoperability for Better Health and Safer Healthcare", SemanticHEALTH Report, 2009, http://ec.europa.eu/information_society/activities/health/docs/publications/2009/2009semantic-health-report.pdf.

[7] K. Milian, A. Bucur and A. ten Teije, "Formalization of clinical trial eligibility criteria: Evaluation of a pattern-based approach", to appear in IEEE BIBM 2012.