Analysis of the Suitability of Existing Medical Ontologies for Building a Scalable Semantic Interoperability Solution Supporting Multi-site Collaboration in Oncology

Ahmed Ibrahim and Anca Bucur
Philips Research Europe
High Tech Campus 34
Eindhoven, the Netherlands

Andre Dekker and M. Scott Marshall
MAASTRO clinic
Dr. Tanslaan 12
Maastricht, the Netherlands

David Perez-Rey and Raul Alonso-Calvo
Universidad Politécnica de Madrid
Boadilla del Monte
Madrid, Spain

Holger Stenzhorn
Saarland University
Building 9
Saarland, Germany

Sheng Yu
University of Oxford
Old Road Campus Research Building
Oxford, United Kingdom

Cyril Krykwinski
Jules Bordet Institute
50, Av. F.Roosevelt
Brussels, Belgium

Anouar Laarif and Keyur Mehta
German Breast Group
Martin-Behaim-Str. 12
Neu-Isenburg, Germany

Abstract—**Semantic interoperability is essential to facilitate efficient collaboration in heterogeneous multi-site healthcare environments. The deployment of a semantic interoperability solution has the potential to enable a wide range of informatics-supported applications in clinical care and research both within a single healthcare organization and in a network of organizations. At the same time, building and deploying a semantic interoperability solution may require significant effort to carry out data transformation and to harmonize the semantics of the information in the different systems. Our approach to semantic interoperability leverages existing healthcare standards and ontologies, focusing first on specific clinical domains and key applications, and gradually expanding the solution when needed. An important objective of this work is to create a semantic link between clinical research and care environments to enable applications such as streamlining the execution of multi-centric clinical trials, including the identification of eligible patients for the trials. This paper presents an analysis of the suitability of several widely-used medical ontologies in the clinical domain: SNOMED-CT, LOINC, MedDRA, to capture the semantics of the clinical trial eligibility criteria, of the clinical trial data (e.g., Clinical Report Forms), and of the corresponding patient record data that would enable the automatic identification of eligible patients. Next to the coverage provided by the ontologies we evaluate and compare the sizes of the sets of relevant concepts and their relative frequency to estimate the cost of data transformation, of building the necessary semantic mappings, and of extending the solution to new domains. This analysis shows that our approach is both feasible and scalable.**

I. INTRODUCTION

Semantic interoperability is essential to facilitate efficient cross-organization collaboration in healthcare by addressing some of the many sources of heterogeneity that currently hamper data and information sharing: (1) Many clinical research and care systems, often home-grown, (2) Structured, semi-structured and free-text documents, (3) Local languages in the patient record systems, English in research systems, (4) Many standards and terminologies, but with relatively low adoption.

The deployment of a semantic interoperability solution has the potential to enable a wide range of informatics-supported applications in clinical care and research, both within a single healthcare organization and in a network of organizations, by enabling exchange of data and information with shared meaning. At the same time, building and deploying a semantic interoperability solution may require significant effort to carry out data transformation and to harmonize the semantics of the information in the different systems.

We aim to implement semantic interoperability in a pragmatic way, relying on existing healthcare standards and ontologies, focusing first on specific clinical domains and key applications, and gradually expanding the solution when needed. The general approach to semantic interoperability is described in [1].

The work described in this paper is carried out in the EURECA[1] project which aims to create a bridge between clinical research and clinical practice by building advanced, scalable and secure solutions which link existing systems such as Clinical Trial (CT) and Electronic Health Record (EHR) systems. Semantic interoperability among EHR and CT systems, consistent with widely-accepted standards and ontologies, lies at the heart of the project.

An important application that we aim to support in the oncology domain is streamlining the execution of multi-centric clinical trials, including the identification of eligible patients for the trials. To support the efficient execution of post-genomic multi-centric clinical trials in cancer, we aim to automatically assess based on the available patient data whether a patient is suitable for enrollment in any of the available clinical trials. The population eligible for a trial is described by a set of free-text eligibility criteria that are both syntactically and semantically complex. The evaluation 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 satisfies these criteria.

In order to facilitate the linkage between clinical trials and patient data, we need to select ontologies which sufficiently capture the content of the eligibility criteria in clinical trials and patients data sets. An important task is to identify candidate subsets of ontologies (instead of the mapping of entire ontologies), which enables us to create scalable and feasible solutions.

This paper presents an analysis of the suitability of several medical ontologies widely-adopted in clinical research: SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms[2]), LOINC (Logical Observation Identifiers Names and Codes[3]), MedDRA (Medical Dictionary for Regulatory Activities[4]), to capture the semantics of the clinical trial eligibility criteria, of the clinical trial data, and of the corresponding patient record data (that would be relevant for the automatic identification of eligible patients). We detect subsets of ontologies that characterize the semantics of the eligibility criteria of trials in various clinical domains in oncology and compare these sets. Next, we evaluate the occurrence frequency of the concepts in the selected oncology domains in order to provide meaningful priorities for the task of mapping these ontology concepts in the eligibility criteria to the patient data model. We further assess the effort required to scale our approach to new domains, measured in terms of additional semantic mappings that need to be developed. Finally, we assess the coverage provided to our domains of interest by the selected ontologies and evaluate the need to extend the sets of selected concepts to best suit our clinical domains.

[1] Enabling information re-Use by linking clinical REsearch and Care. http://eurecaproject.eu/

[2] http://www.ihtsdo.org/snomed-ct/

[3] http://www.loinc.org/

[4] http://www.meddra.org/

This work focuses on the definition of the core dataset (defined as in [1]) that sufficiently covers the semantics of our clinical domains of interest. The identification of the core dataset relies on: (1) an automatic identification of the concepts sets in the eligibility criteria of clinical trials in the domains of breast cancer, lung cancer, sarcoma and nephroblastoma, (2) on the identification of the sets of concepts of selected ontologies that appear in the available patient record datasets, and (3) an evaluation of both sets of ontology concepts by the clinical and knowledge experts and an assessment of coverage of the selected ontologies for the domains of interest.

In [2] we carried out an analysis in identifying the semantics of clinical trial eligibility criteria with the same set of ontologies: SNOMED-CT, LOINC and MedDRA. We identified the subsets of the ontologies (since mapping entire ontologies is not feasible because of their respective size) that sufficiently capture the content of the eligibility criteria of trials in breast cancer and compared with trials in cancers other than breast and in the cardiovascular domain. We also evaluated whether our modular approach for the selection of the sets of concepts based on the clinical domains is scalable and feasible, and prioritized relevant concepts based on their frequency in the breast cancer subset and on their co-occurrence in trials in the other clinical domains. 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, and that the reuse of concepts across trials is very significant. This conclusion is strengthened by the current work. In addition, the current work also analyzes patient-record datasets and provides an evaluation of the coverage of the ontologies for each clinical domain of interest. This evaluation was carried out by clinical and knowledge experts.

II. EXTENDING THE SEMANTIC SOLUTION

The first implementation of our semantic solution for the domain of breast cancer is described in [3] and in [4] we present the initial results towards developing mappings between the eligibility criteria and the patient data model based on the HL7-RIM standard and on the selected ontologies.

In this paper we continue to further extend and assess our semantic approach in other clinical domains in oncology: lung cancer, sarcoma and nephroblastoma, and with new datasets in all four oncology domains of focus. We identify the subsets of the ontologies that sufficiently capture the content of the eligibility criteria of trials and of the relevant data collected in the patient records in the clinical domains of lung cancer, sarcoma and nephroblastoma and compare with our results in breast cancer updated with the new datasets. From the patient record datasets we extracted and analyzed the concepts that are relevant for the oncology patient management. We also identified relevant concepts that are currently not supported by our semantic solution in order to evaluate the extensibility of the solution and the needed effort.

The analysis of the semantics of the datasets in the oncology domains of interest is based on three medical ontologies: SNOMED-CT, MedDRA, and LOINC. These ontologies are considered the best choices due to their wider adoption in both clinical research and care. The scalability of our solution is achieved by modularization i.e., we identify a core subset of SNOMED-CT that covers each clinical domain of interest and allows us to model the available datasets. We continue this process when we need to add new concepts related to new trials and data sources in an already captured clinical domain, and we define new modules for each new domain.

III. DESCRIPTION OF THE EXPERIMENTS

The experiments consist of the identification of the core dataset information in (a) clinical trial description, such as clinical trial eligibility criteria, and (b) care datasets provided by clinical sites (e.g., from the EHR system).

A. The identification of the core dataset information in clinical trial description

For the analysis of the semantics of the clinical trial eligibility criteria, we selected trials published on ClinicalTrial.gov and extracted the eligibility criteria of those trials from the following clinical domains: breast cancer, lung cancer, sarcoma and nephroblastoma. ClinicalTrials.gov is a service of the U.S. National Institute of Health and lists more than 157,327 trials with locations in all 50 states and in 185 countries[5]. TABLE 1 shows the number of trials in each of the four domains.

TABLE 1 NUMBER OF TRIALS IN THE EVALUATION

Clinical domain			
breast cancer	***Lung cancer***	***Sarcoma***	***Nephroblastoma***
4232	1598	421	172

In the following section, we investigate the semantic similarity among clinical trials in the selected clinical domains, and compare the sets of concepts for these domains and the selected ontologies. Our main focus is finding semantic overlap between breast cancer and the other domains as this enables us to reuse a part of our semantic solution (which initially focuses on breast cancer). For the analysis we extract the concepts that are found in the eligibility criteria and use an annotator from BioPortal to identify the ontology concepts present in those criteria. BioPortal is an open repository of biomedical ontologies that provides access via Web browsers and Web services to ontologies [5]. The BioPortal results include information such as identifiers, labels and the UMLS semantic type of the concepts. The semantic types can provide additional information about the semantics of the criteria and identify similarity between concepts. The following ontologies were added to the annotator: SNOMED-CT, MedDRA and LOINC.

[5] http://www.clinicaltrials.gov/

B. The identification of the core dataset information in available datasets provided by clinical partners

The clinical users involved in the EURECA project and contributing to this work are the University of Oxford (UOXF), Institute Jules Bordet (IJB), MAASTRO clinic, German Breast Group (GBG), and the University of Saarland (UdS). The datasets contain EHR data, Case Report Forms (CRF) used in trials such TOP Trial and SIOP 2001/GPOH, Neo-adjuvant (GeparQuattro), Adjuvant (GAIN) and metastatic (TBP) breast cancer trials, and cancer registry concepts (such as morphology and topography) that have been manually extracted. The diseases in which datasets are provided are: breast cancer, radiation oncology in breast cancer and lung cancer, sarcoma and nephroblastoma. In this context, for data transformation we needed to address many of the relevant types of heterogeneity (e.g. system, structure, language, terminology, domain, etc.).

For instance for the MAASTRO clinic dataset we extracted and analyzed EHR data which was initially represented as a mix of structured fields and free-text in the Dutch language. The data processing steps included the identification and extraction of the relevant concepts, the identification of corresponding ontology concepts when available and an evaluation of the mappings with the clinical experts.

The clinical domain of MAASTRO is radiation oncology, and we particularly chose this dataset for the evaluation because it contains a mixture of free-text and structured data. The data files are annotated with NCIt (National Cancer Institute Thesaurus) codes. The free-text includes oncological history, medication and medical history. The NCIt annotation contains information such as the name of the disease, the date of diagnosis and the TNM stage [6].

IV. EVALUATION RESULTS

In this section we present (1) an analysis of the semantics of the clinical trial eligibility criteria based on relevant medical ontologies: SNOMED-CT, MedDRA and LOINC, (2) evaluate the sets of concepts that appear in a patient dataset provided by a clinical partner, and (3) an evaluation of the results by clinical partners.

A. The identification of the core dataset information in clinical trial description

TABLE 2 compares the sets of concepts for the four clinical domains and the three ontologies selected. We denote **L** as the set of lung cancer concepts, **B** as the set of breast cancer concepts and **SN** as the union of the sarcoma and nephroblastoma concepts (since the sets for these domains are relatively small compared to the lung cancer and breast cancer corpuses). The tables show that the lung cancer and breast cancer corpuses have significant subsets that are specific for those diseases, but also a large overlap. This implies that a large amount of concepts currently used in our semantic layer for breast cancer will be also relevant for lung cancer. We can also observe that for LOINC the largest set is the one that is the overlap among the four domains. For SNOMED-CT we observe that this is the second largest set.

TABLE 2 SETS OF CONCEPTS FOR LUNG CANCER (L), SARCOMA AND NEPHROBLASTOMA (SN) AND BREAST CANCER (B)

Subset	SNOMED-CT	MedDRA	LOINC
L – (B ∪ SN)	1666	527	427
L ∩ B	1976	499	579
L ∩ SN	34	13	61
L ∩ B ∩ SN	2225	399	1006
SN – (B ∪ L)	506	164	120
B ∩ SN	76	18	33
B – (SN ∪ L)	3392	840	727

FIGURE 1, FIGURE 2 and FIGURE 3 show for the three corpuses of trials the distribution of concepts across trials (for SNOMED-CT). We only show the top most frequent concepts and each concept is counted once per trial. The figures show that a large ratio of criteria is similar, but that new trials may introduce new concepts. A relatively small group of concepts occur in a large number of trials and there are concepts that occur less frequent for specific trials. We also see this phenomenon in the results presented in [2].

FIGURE 1 the number of lung cancer (y-axis) trials that include the top 500 most frequently occurring SNOMED-CT concepts (x-axis). Concepts were counted once per trial.

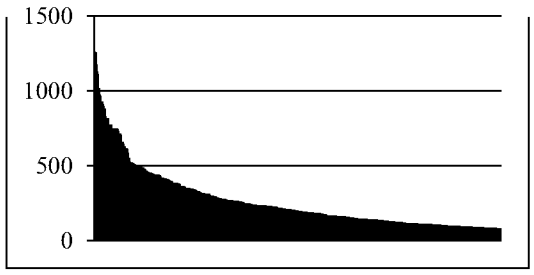


FIGURE 2 the number of sarcoma (y-axis) trials that include the top 500 most frequently occurring SNOMED-CT concepts (x-axis). Concepts were counted once per trial.

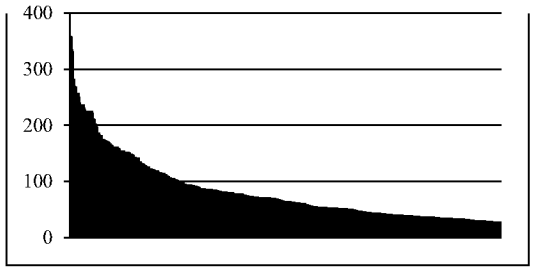


FIGURE 3 the number of nephroblastoma (y-axis) trials that include the top 500 most frequently occurring SNOMED-CT concepts (x-axis). Concepts were counted once per trial.

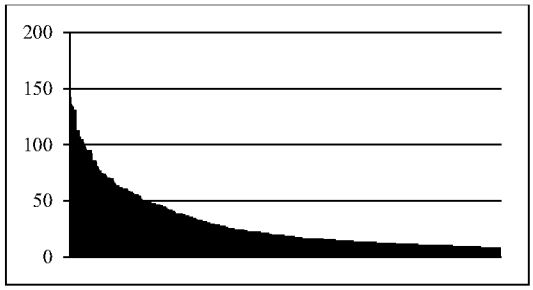


TABLE 3 shows the averages of the number of ontology concepts per trial in the domains lung cancer, sarcoma and nephroblastoma, for the three selected ontologies.

TABLE 3 THE AVERAGE NUMBER OF TRIALS PER ONTOLOGY

	SNOMED-CT	*MedDRA*	*LOINC*
lung cancer	192	25	112
sarcoma	245	31	140
nephroblastoma	206	25	115

B. The identification of the core dataset information in a breast cancer dataset provided by a clinical site

In this section we evaluate the concepts extracted from the MAASTRO breast cancer dataset. The groups of files are NCIt encoded, and free-text oncological history, medical history and medication. For each group, we extract the concepts that are found in the data and use the Bioportal annotator to identify the ontology concepts.

1) NCI Thesaurus

Every patient record in the MAASTRO dataset has at least one NCIt encoded element. An element has attributes such as an id, a code, and a preferred name. TABLE 4 shows a list of the diseases (with associated NCIt codes) and the coverage of the ontologies, for the breast cancer dataset. We used mappings that were created by the LOOM algorithm of NCBO[6] and a clinical expert at MAASTO to validate and to complement the mappings.

[6]http://bioontology.stanford.edu/wiki/index.php?title=BioPortal_Mappings&redirect=no

TABLE 4 COVERAGE BY MEDICAL ONTOLOGIES OF CONCEPTS IN MAASTRO BREAST CANCER DATASET (NO COVERAGE=0, 100% COVERAGE=1.0)

	Mapped To		
Breast cancer dataset (NCIt code)	***SNOMED-CT***	***MedDRA***	***LOINC***
Breast Neoplasm (C2910)	X	X	X
Invasive Ductal Carcinoma, Not Otherwise Specified (C4194)	X	X	
Invasive Lobular Breast Carcinoma (C7950)	X	X	
Ductal Breast Carcinoma In Situ (C2924)	X	X	X
Malignant Breast Neoplasm(C9335)	X	X	
Medullary Breast Carcinoma (C9119)		X	
Mixed Lobular and Ductal Breast Carcinoma (C5160)	X		
Tubular Breast Carcinoma (C9135)		X	
Adenoid Cystic Breast Carcinoma (C5130)		X	
Malignant Breast Phyllodes Tumor(C4504)	X	X	
Coverage	**0.70**	**0.90**	**0.20**

2) Oncological history

The free-text oncological history contains TNM cancer staging codes and the dates associated with them. The TNM staging system is one of the most widely used cancer staging systems. The TNM system is based on the size and/or extent (reach) of the primary tumor (T), the amount of spread to nearby lymph nodes (N), and the presence of metastasis (M) or secondary tumors formed by the spread of cancer cells to other parts of the body [6]. We use regular expressions to extract the TNM stages and incorporate a pattern such that we can also extract the date of occurrence, e.g., the Dutch free-text "...9 November 2012 pT1N0M0 mammacarcinoom…" contains "(pT1N0M0, 2012)". TABLE 5 shows a list of the most frequent TNM cancer staging codes that were extracted.

TABLE 5 MOST FREQUENT OCCURING TNM PARAMETERS WITH CORRESPONDING DEFINITION

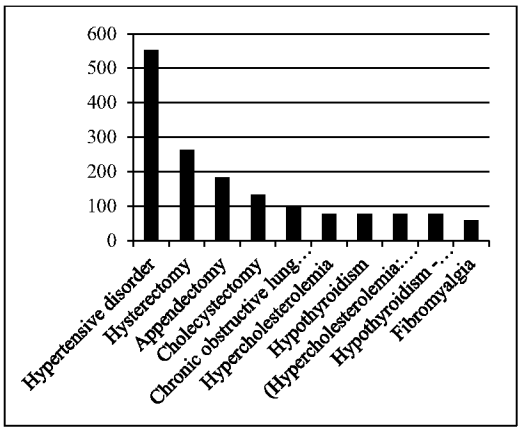
Category	***Definition***[7]	***Identified by***
cTis	Carcinoma in situ	SNOMED-CT
cT1 / pT1	Tumor 2 cm or less in greatest dimension	SNOMED-CT
cT2	Tumor more than 2 cm but not more than 5 cm in greatest dimension	SNOMED-CT
N0	No regional lymph node metastasis	SNOMED-CT
N1	Spread to movable ipsilateral axillary lymph node(s)	SNOMED-CT
M0	No distant metastasis	SNOMED-CT

[7] The list of definitions has been retrieved from [6].

3) Medical history

FIGURE 4 shows the number of occurrences of the top 10 most frequent annotated concepts extracted from the medical history section of the breast cancer dataset. The figure shows that 'Hypertensive disorder', which can be caused by several breast cancer treatments, and 'Hysterectomy', which can lower the risk of certain cancers, are frequently occurring concepts within the dataset. In some cases we noticed that concepts with the same label (but with different identifiers) can also appear in different branches in the SNOMED-CT hierarchy (i.e., a term can have multiple interpretations). We needed clinical experts and the UMLS semantic types for further disambiguation.

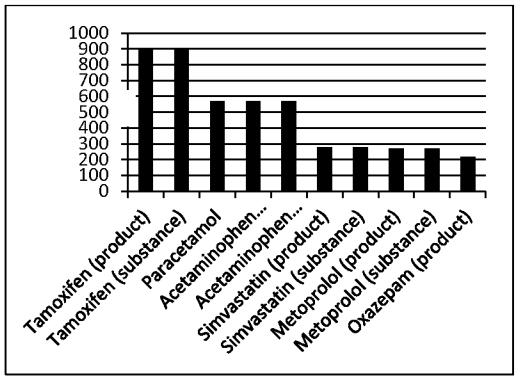
FIGURE 4 the number (y-axis) of occurrences of the top 10 most annotated concepts extracted from the medical history section (x-axis) in SNOMED-CT.



4) Medication

FIGURE 5 shows the number of occurrences of the top 10 most frequent annotated concepts extracted from the medication section of the breast cancer dataset. The figure shows that 'Tamoxifen' (a drug that is used to treat hormone receptor-positive breast cancer) is the most frequently occurring concept found in the breast cancer dataset. The annotator identified this concept as a 'Pharmaceutical/Biological product' and as a 'Substance'.

FIGURE 5 the number (y-axis) of occurrences of the top 10 most annotated concepts extracted from the medication section (x-axis) in SNOMED-CT. The concepts 'Tamoxifen', 'Acetaminophen', 'Simvastatin' and 'Metoprolol' are identified as a product and as a substance in the SNOMED-CT hierarchy.



5) Semantic types

TABLE 6 shows the semantic types of the NCIt encoded section of the breast cancer dataset. The table shows that the semantic type 'Neoplastic Process', which in turn is a 'Disease or Syndrome' in the UMLS, is the most frequent occurring type. We can also see the same findings regarding the free-text medication (TABLE 7). Almost 90% of the data in that section is identified as a 'Pharmacologic Substance'. In the medical history section (TABLE 8) we see that 66% is identified as a 'Disease or Syndrome'.

TABLE 6 RATIO OF THE SEMANTIC TYPES FOR NCIt ENCODED SECTION OF THE BREAST CANCER DATASET (RELATIVE TO TOTAL NUMBER OF CONCEPTS)

Semantic type	*Ratio*
Neoplastic Process	0.559
Body Part, Organ, or Organ Component	0.223
Qualitative Concept	0.204
Spatial Concept	0.011
Disease or Syndrome	0.010
Finding	0.001

TABLE 7 RATIO OF THE MOST FREQUENT SEMANTIC TYPES FOR THE MEDICATION SECTION OF THE BREAST CANCER DATASET (RELATIVE TO TOTAL NUMBER OF CONCEPTS)

Semantic type	*Ratio*
Pharmacologic Substance	0.873
Organic Chemical	0.794
Disease or Syndrome	0.037
Laboratory Procedure	0.034
Steroid	0.033
Element, Ion, or Isotope	0.028
Biologically Active Substance	0.024
Hormone	0.022
Amino Acid, Peptide, or Protein	0.020
Vitamin	0.014

TABLE 8 RATIO OF THE MOST FREQUENT SEMANTIC TYPES FOR THE MEDICAL HISTORY SECTION OF THE BREAST CANCER DATASET (RELATIVE TO TOTAL NUMBER OF CONCEPTS)

Semantic type	*Ratio*
Disease or Syndrome	0.663
Therapeutic or Preventive Procedure	0.235
Finding	0.055
Pathologic Function	0.039
Body Part, Organ, or Organ Component	0.027
Congenital Abnormality	0.026
Anatomical Abnormality	0.025
Qualitative Concept	0.011
Clinical Attribute	0.006
Acquired Abnormality	0.004

C. The coverage evaluation carried out by clinical experts

TABLE 9 shows the coverage of the selected ontologies for each of the dataset introduced in Section III.B. The table indicates that SNOMED-CT has the highest coverage rate. Further analysis, however, revealed that SNOMED-CT is insufficient to cover more specific concepts such as radiotherapy concepts and gene related concepts (e.g., HER2).

TABLE 9 COVERAGE OF DIFFERENT TERMINOLOGIES (NO COVERAGE=0, 100% COVERAGE=1.0)

Dataset		*SNOMED-CT*	*MedDRA*	*LOINC*
UOXF	BC dataset	0.75	0.33	0.37
	Sarcoma dataset	0.74	0.55	0.37
IJB	CRF data TOP Trial	0.91	0.43	0.66
	Cancer Registry Morphology	0.98	0.35	0.01
	Cancer Registry Topography	0 .05	0.01	0.007
MAASTRO	EHR	0.70	0.90	0.20
GBG	TBP	0.99	0.70	0.29
	GAIN	0.95	0.62	0.22
	GeparQuattro	0.95	0.61	0.21
UdS	SIOP	0.96	0	0.04

V. RELATED WORK

In this paper we present an analysis of the semantics of the eligibility criteria and datasets of oncology trials and of the relevant patient record data based on widely-adopted medical ontologies: SNOMED-CT, MedDRA and LOINC.

SNOMED-CT 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 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 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.

In our previous work in the INTEGRATE[8] project we analyzed the semantics of the eligibility criteria of clinical trials in the INTEGRATE domain of interest which is breast cancer. We defined the core dataset at the center of the semantic solution, linking trial descriptions to patient data as: "*Soundly defined and agreed-upon clinical structures consisting of stand-based concepts, their relationships, quantification etc., that together sufficiently describe the clinical domain*" [1]. In this paper we leverage the work in the INTEGRATE project and extend our semantic approach to other clinical domains such as lung cancer, sarcoma and nephroblastoma.

In [7] the authors present a method that uses the Web Ontology Language (OWL) and the Semantic Web Rule Language (SWRL) for automatic recruitment of a patient to available clinical trials. The aim of their work is to show how it is possible to represent eligibility criteria of clinical trials using SWRL on top of a large domain specific ontology: NCI thesaurus. Their evaluations results indicate that this ontology provides the best coverage of the terms that appear in the patient datasets and in the selected eligibility criteria used in the set-up. Patient datasets and eligibility criteria, in the domain of prostate cancer, are represented with the Web Ontology Language and SWRL. The SWRL rules are queried over the observations containing patients' data to verify which inclusion and/or exclusion criteria are met. Then, the rules are computed and executed with the Jess rule engine. The end result is a list of inclusion and exclusion criteria with their associated observations. The authors conclude that a large majority of the criteria of selected trials can be represented but that a lack of corresponding concepts in NCIt is the main cause of failure.

In [8] the authors present a modeling strategy of eligibility criteria in OWL that leverages open world assumption to address the missing information problem. The proposed OWL design pattern deals with scenarios in which inclusion criteria cannot be proven and exclusion criteria cannot be proven to be false. The analysis of a clinical trial shows that ignoring missing information (i.e., eligibility criteria which cannot be assessed) leads to too many rejections. Their strategy identified 30 patients, which were initially rejected, as potentially eligible for a clinical trial.

In [9] the authors present a feasibility study for an ontology-based approach to match patient records to clinical trials. The process involves formulating trials as queries and using the SHER reasoner to match the queries against a knowledge base to retrieve eligible patients. The authors present results in which challenges such as identifying and eliminating noise in patient data, and dealing with incomplete patient information are solved using their approach, for a real world patient dataset that is used in the analysis.

[8] Driving excellence in Integrative Cancer Research through Innovative Biomedical Infrastructures. http://www.fp7-integrate.eu/

VI. DISCUSSION AND CONCLUSIONS

In our previous work we proposed a flexible and scalable approach to semantic interoperability and provided a first implementation. Our approach relies on the selection of suitable and widely-adopted ontologies that can provide sufficient coverage for the clinical domains and applications of interest. These ontologies are used to define the core dataset, i.e., subsets/modules that are used in our solution to provide semantic links among datasets in various relevant systems.

For this modular approach to be feasible and scalable, (1) the modules capturing the semantics of each domain need to be relatively small to limit the effort required for data transformation and for the development of mappings, (2) the extensions to new domains or applications need to be manageable in terms of the numbers of concepts that need to be added, and (3) the selected ontologies need to provide good coverage (i.e., capture the semantics) for the relevant domains and applications.

In [2] we have evaluated the feasibility of this approach for breast cancer in the context of supporting clinical trials. In this paper we assess the feasibility and scalability of this solution by extending it to new clinical domains and evaluating additional datasets from clinical care and research.

We focus on the definition and analysis of the core dataset (i.e., modular subsets of existing ontologies) that sufficiently cover the semantics of the clinical domains of interest. The identification of the core dataset relies on: (1) an automatic evaluation of the concepts sets in the eligibility criteria of clinical trials in the domains of breast cancer, lung cancer, sarcoma and nephroblastoma, and (2) on the evaluation of the sets of concepts of selected ontologies that appeared in patient datasets (both from research and care) provided by clinical sites.

Our results indicate that a relatively small set of concepts occurred in a large number of clinical trials and that there were concepts that occurred less frequently (only in specific trials). With these findings we can prioritize the implementation of semantic mappings starting with the most frequent concepts. Also, our results show that the effort of adding new trials is low since the additional sets of concepts that need to be mapped to relevant data are small.

It can also be observed that the semantic overlap between the lung cancer and breast cancer domains is large. This implies that we can reuse a large part of our semantic mappings developed for breast cancer when extending the solution to the lung cancer domain. It also shows that this approach can be further generalized to new clinical and application domains, and that the cost of extending the implementation is relatively low.

The evaluation of the different datasets by the clinical experts indicated that SNOMED-CT provides the best coverage for our data. However, further analysis revealed that SNOMED-CT is insufficient to cover more specific concepts related for instance to radiotherapy or genomic data. For these domains other ontologies need to be evaluated.

ACKNOWLEDGEMENT

This work has been partially funded by the European Commission through the EURECA project (FP7-ICT-2011-288048)

REFERENCES

[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] A. Bucur, J. v. Leeuwen, D. Perez-Rey, R. C. Alonso, B. Claerhout and K. . d. Schepper, "Identifying the Semantics of Eligibility Criteria of Clinical Trials based on relevant Medical Ontologies," in *Bioinformatics & Bioengineering (BIBE), IEEE 12th International Conference*, Larnaca, 2012.

[3] S. Paraiso-Medina and et al., "Semantic Interoperability Solution for Multicentric Breast Cancer Trials at the Integrate EU Project," in *6th International Conference on Health Informatics, pp. 34–42*, 2013.

[4] A. Rico-Diez and et al., "SNOMED CT Normal Form and HL7 RIM binding to normalize clinical data from cancer trials," in *IEEE 13th International Conference on Bioinformatics & Bioengineering (BIBE)*, 2013.

[5] N. Noy, N. Shah, P. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. Rubin, M. Storey, C. Chute and M. Musen, "BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse.," *Nucleic Acids Res.,* 2009 (37)W170-3.

[6] F. Greene, AJCC cancer staging handbook. American Joint Committee on Cancer, New York: Springer, 2002.

[7] P. Besana, M. Cuggia, O. Zekri, A. Bourde and A. Burgun, "Using Semantic Web Technologies for Clinical Trial Recruitment," *The Semantic Web – ISWC 2010,* vol. 6497, pp. 34-49, 2010.

[8] O. Dameron, P. Besana, O. Zekri, A. Bourde, A. Burgun and M. Cuggia, "OWL Model of Clinical Trial Eligibility Criteria Compatible With Partially-known Information," *Journal of Biomedical Semantics,* 2013.

[9] P. Chintan, J. Cimino, J. Dolby, A. Fokoue, A. Kalyanpur, A. Kershenbaum, L. Ma, E. Schonberg and K. Srinivas, "Matching Patient Records to Clinical Trials Using Ontologies," *The Semantic Web-Lecture Notes in Computer Science,* vol. 4825, pp. 816-829, 2007.