

# A Semantic Interoperability Approach to Support Integration of Gene Expression and Clinical Data in breast cancer

Raul Alonso-Calvo<sup>1</sup>, Sergio Paraiso-Medina<sup>1</sup>, David Perez-Rey<sup>1</sup>, Enrique Alonso-Oset<sup>1</sup>, Ruud van Stiphout<sup>2</sup>, Sheng Yu<sup>2</sup>, Marian Taylor<sup>2</sup>, Francesca Buffa<sup>2</sup>, Carlos Fernandez-Lozano<sup>3</sup>, Alejandro Pazos<sup>3</sup> and Victor Maojo<sup>1</sup>

<sup>1</sup> Biomedical Informatics Group, DIA & DLSIIS, ETSI Informáticos, Universidad Politécnica de Madrid, Spain; (ralonso, sparaiso, dperez, enriquealonso, vmaojoo)@infomed.dia.fi.upm.es

<sup>2</sup> Department of Oncology, Old Road Campus Research Building, Oxford OX3 7DQ, United Kingdom; (ruud.vanstiphout, sheng.yu, marian.taylor, francesca.buffa)@oncology.ox.ac.uk

<sup>3</sup> Department of Information and Communication Technologies, Faculty of Computer Science, University of A Coruña, 15071 A Coruña, Spain; (carlos.fernandez, apazos)@udc.es

## Corresponding author:

Raúl Alonso Calvo

e-mail: ralonso@infomed.dia.fi.upm.es; tel.: +34-91-336-74-67

DLSIIS, ETSI Informáticos, Universidad Politécnica de Madrid

Campus de Montegancedo S/N, 28660, Boadilla del Monte, Spain

## Abstract:

### Introduction

The increment of information in healthcare due to the introduction of omics data and advances in technologies involved in clinical treatment have led to a broad range of approaches to represent clinical information. Within this context, patient stratification across health institutions due to omic profiling presents a complex scenario to carry out multi-center clinical trials.

### Methods

This paper presents a standards-based approach to ensure semantic integration required to facilitate the analysis of clinico-genomic clinical trials. To ensure interoperability across

different institutions, we have developed a Semantic Interoperability Layer (SIL) to facilitate homogeneous access to clinical and genetic information, based on different well-established biomedical standards and following International Health (IHE) recommendations.

## **Results**

The SIL has shown suitability for integrating biomedical knowledge and technologies to match the latest clinical advances in healthcare and the use of genomic information. This genomic data integration in the SIL has been tested with a diagnostic classifier tool that takes advantage of harmonized multi-center clinico-genomic data for training statistical predictive models.

## **Conclusions**

The SIL has been adopted in national and international research initiatives, such as the EURECA-EU research project and the CIMED collaborative Spanish project, where the proposed solution has been applied and evaluated by clinical experts focused on clinico-genomic studies.

## **Keywords:**

Clinical Research Informatics, Semantic Interoperability, Data Integration, Diagnostic Classifier, Gene Expressions, Biomedical Terminologies

---

## **1. Introduction**

Clinical trial complexity is dramatically increasing as new genetic and molecular variables are gathered in clinical settings [1]. Due to the costs of such clinical studies and challenges for recruiting trial cohorts, they often involve multiple clinical institutions [2]. New data management methods are therefore required by clinical users and investigators from institutions involved in multi-center clinical research [3]. In most cases, researchers need to know the different data representations of the institutions participating in the study and significant manual data management is required [4]. To facilitate certain processes required to achieve semantic integration from heterogeneous sources in the area (e.g., clinical trial management systems, electronic health records or laboratory systems, among others) (semi-) automatic methods have been recently addressed by international initiatives [5].

Several efforts have recently focused on facilitating communication and exchange of information between clinical systems by using biomedical standards [6]. In general,

interoperability initiatives provide an underlying data model for different areas. Examples of these initiatives are, to mention a few relevant examples, the Observational Medical Outcomes Partnership (OMOP) [7], Integrating Biology and the Bedside (i2b2) [8], the HL7 Reference Information Model (RIM) [9], Fast Healthcare Interoperability Resources (FHIR) [10], Integrating the Healthcare Enterprise (IHE) [11] or PCORnet [12]. These initiatives have been developed with the objective of obtaining valuable results in the clinical research area, but few of them have actually exploited the benefits of the analysis and interaction with genetic information and related terminologies.

Clinical terminologies have been historically used in medicine to classify and categorize diseases. One of the most relevant terminologies is SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms) [13]. SNOMED-CT is a general purpose clinical vocabulary distributed by The International Health Terminology Standards Development Organization (IHTSDO), with over 400 thousands concepts, 1 million of descriptors and more than 1 million of relationships between them. While SNOMED-CT provides broad coverage, there are other terminologies oriented to more specific clinical areas. Logical Observation Identifiers Names and Codes (LOINC) [14], developed by the Regenstrief Institute in Indiana, USA, is a clinical terminology for identifying laboratory and clinical test results.

In the context of breast cancer research, recent studies show that more than 5% of breast cancer patients might be hereditary [15], caused by gene information inherited from their families' relatives. "All-purpose" terminologies such as SNOMED-CT frequently do not provide the highest coverage for this specific domain. In this area, terminologies such as the HUGO Gene Nomenclature Committee (HGNC) [16] contain only genetic concepts. HUGO is an international classification of the human gene nomenclature, and an open access database containing more than 33,000 gene names and symbols at the time of writing. The majority of these items are protein-coding genes, but they also contain pseudogenes, non-coding RNAs, phenotypes and genomic features.

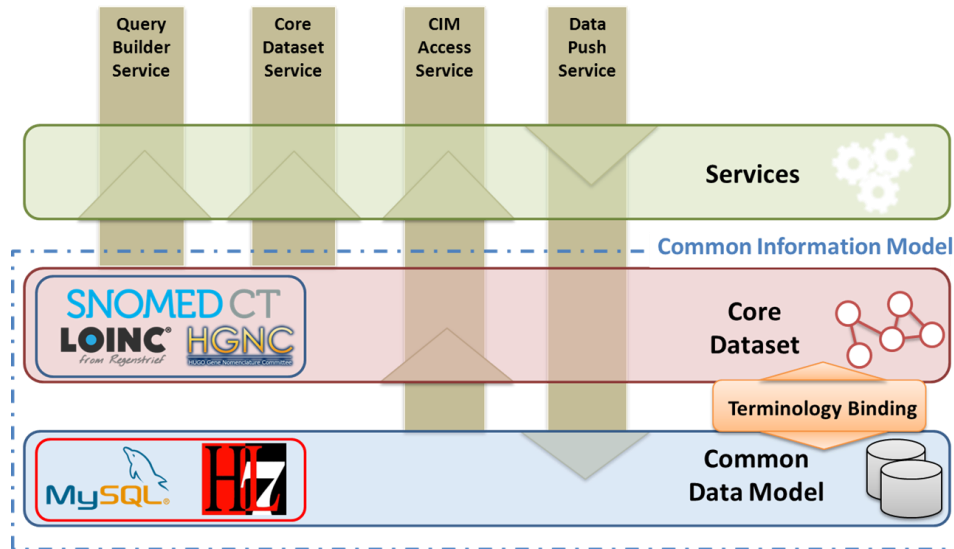
With an increasing focus on genomics, in last years the number of translational biomedicine solutions has significantly increased. Different approaches intend to exploit the availability of omic data correlated with clinical data to enhance prevention, diagnosis, and therapies [17][18]. Standardization initiatives in biomedicine such as transSMART [19], HL7 in standard v3 [20], HL7 FHIR [21] and CDISC [22] are actively working in translational biomedicine. I.e. CDISC has delivered the Study Data Tabulation Model (SDTM) [23] for representing the clinical domain; CDISC also propose an implementation guide for pharmacogenomics and pharmacogenetics (SDTMIG-PGx) [24], defining relations of

biospecimen and genetics-related data. Research projects such as the cancer translational research informatics platform (caTrip) [25] or BioShare [26], have proposed the exploitation of BioBank data together with EHR data on breast cancer, providing insights on the viability of implementing translational platforms. Other networks of as the Electronic Medical Records and Genomics Network (eMERGE) have been also created to explore different ways of integrating omic and clinical data as summarized in [27].

We describe our proposed Semantic Interoperability Layer (SIL), and selected examples of its applications within international research projects: EURECA (Enabling information re-use by linking clinical Research and Care) [28] and CIMED (Collaborative Project on Medical Informatics). The objective is to investigate if such standards-based approach can be used to integrate genomic information and support the analysis of interactions with clinical genetic information in breast cancer studies and diagnostic classifier analysis. This Semantic Interoperability Layer uses standard terminologies as a vehicle for addressing two main challenges in multi-centric interoperability: harmonizing heterogeneities from different data sources as well as for integrating omic and clinical data.

## **2. Materials and Methods**

To homogenize common information across different clinical settings, such as clinical trial management (CTMS) systems, electronic health records (EHR), laboratory information management systems (LIMS) and others, in this work we propose a standard-based SIL including one common information model (CIM) and a set of services as homogenous endpoints to access data. As shown in **Figure 1**, the proposed SIL is defined by the interaction between the CIM and services for data access. The CIM is composed of three main components: i) the common data model (CDM), ii) the core dataset (terminologies) and (iii) the linking between them (terminology binding). The SIL was designed as the basis for software services and tools developed within the project, which are focused on enhancing clinical research with genetic information.



**Figure 1:** Interaction diagram of SIL components

To analyze the interaction of breast cancer gene expressions with clinical data, a set of services for data retrieval were defined within the SIL. These services provided uniform access to data stored in the SIL, exploiting semantic and abstraction capabilities of the CIM. The core dataset integrates terminologies such as SNOMED-CT, HGNC and LOINC for covering the clinical scenario domain [29]. The CDM is a HL7 RIM-based structure required to homogenize data models of information systems from different institutions. Finally, a binding solution for linking the concepts from clinical terminologies to the corresponding CDM classes has been developed. In this work, the SIL provides a standard infrastructure to integrate clinical and genetic information exploited by diagnostic classifiers.

## 2.1. Common information model support for genetic information

### 2.1.1 Core Dataset extension with genetic information

In general, biomedical data integration requires the use of terminologies to annotate data sources and facilitate the integration of data from heterogeneous sources. After the analysis of data sources of different projects, the approach adopted to create the core dataset was to select complementary subsets of widely used terminologies and ontologies. The core dataset is mainly based on SNOMED-CT but it was extended with other domain specific terminologies, such as LOINC for laboratory tests and HGNC for gene names. Selected terminologies have been integrated together into a Web Ontology Language (OWL) file and loaded into a Sesame [30] Server. The Core Dataset is therefore available through a common SPARQL endpoint for the rest of the SIL components.

When harmonizing data with terms coded in other terminologies, such as ICD [31], Gene Ontology [32] and the NCI thesaurus [33], original concepts are annotated with Core Dataset concepts. However, selected concepts for annotating original data to core dataset may lack certain information contained in the original concept. For this reason, the semantic layer can store both the original and the annotated code. The annotation process strongly depends on each data source, from original data sources structured with coded values where automatic translations using UMLS are possible, to free-text data sources where annotation is a more laborious process that could be enhanced using NLP techniques [29]. In any case, after the annotation process, a manual validation of mapped terms by a domain expert is required.

The main goal of the Core Dataset is to provide a comprehensive terminology to cover the source data. SNOMED-CT covered nearly all clinical concepts from the EURECA project, but lacked specific concepts related to radiotherapy and genes. For allowing the mapping of those concepts not covered, LOINC and the HGNC are used in combination with SNOMED-CT.

#### 2.1.2 Common Data Model (CDM)

The CDM is the structure responsible of representing and storing data from clinical institutions. To facilitate the integration of legacy systems that may use built-in or open source, a relational database based on HL7 Reference Information Model (RIM) was developed. Different technologies were analyzed to implement the common data model [34], selecting a relational database due to performance for large patient datasets. To support semantic querying through SPARQL, morph R2RML [35] was used for mapping common data model implemented relational schema to a virtual ontology. SPARQL can be then executed to retrieve the required data from the underlying relational database.

#### 2.1.3 Terminology binding of clinical and genetic information

The Terminology Binding component defines the mapping between Core Dataset concepts and the CDM. It defines in which fields of the CDM, each concept from the core dataset could be stored [36].

While HL7 RIM is able to represent a wide domain of clinical information, it also allows storing the same piece of information in different ways. As the same concept may have different contexts, they can be stored in different CDM attributes. For this reason, a core dataset automatic normalization process to homogenize the representation of these concepts was developed. This normalization process uses SNOMED Normal Form [37] for decomposing

complex concepts into a combination of atomic concepts. Storing only normalized SNOMED-CT concepts in the data model simplifies the binding between concepts and the RIM following IHE recommendations [38].

Binding information has been attached to each Core Dataset concept, linking to its corresponding common data model attribute by including annotations in the Core Dataset OWL file. For this purpose IHE and the TermInfo project recommendations have been used, which associate some concepts of SNOMED-CT to an HL7 RIM class. The RIM class association is propagated, from SNOMED-CT concepts where it is defined, to all their subconcepts, labeling them with the same RIM class. Other terminologies included in Core Dataset are also linked to RIM classes and attributes, but in this case the linking is easier than in SNOMED-CT case. Concretely, all LOINC concepts correspond to code attribute of HL7 RIM Observation; and gene names in HGNC has been bound to code attribute of the HL7 RIM Entity class.

## *2.2. Services for data load & retrieval*

As depicted in *Figure 1*, a set of services for accessing the CDM and core dataset knowledge were developed. The SIL solution is based on SOAP architecture through HTTPS communication protocol ensuring the security of the services. These services allow the abstraction of the CDM schema representation relying only in the Core Dataset concepts and a clinical context. Using the Query Builder Service it is possible to obtain necessary information to build a query of any concept from the core dataset, by providing only that concept. The Core Dataset service contains a set of methods to query the semantic repository of the terminology and to normalize core dataset expressions into a more generic concepts.

A standardized service for querying and populating the CIM is required to seamlessly retrieve data across applications. In order to provide semantic reasoning methods supported by the core dataset, a CIM Access service and a Data Push service were also developed. The CIM Access Service provides an SPARQL query endpoint to retrieve HL7 RIM-based data. It allows use of hierarchical information over normalized data stored in the CDM by using Core Dataset relationships [39].

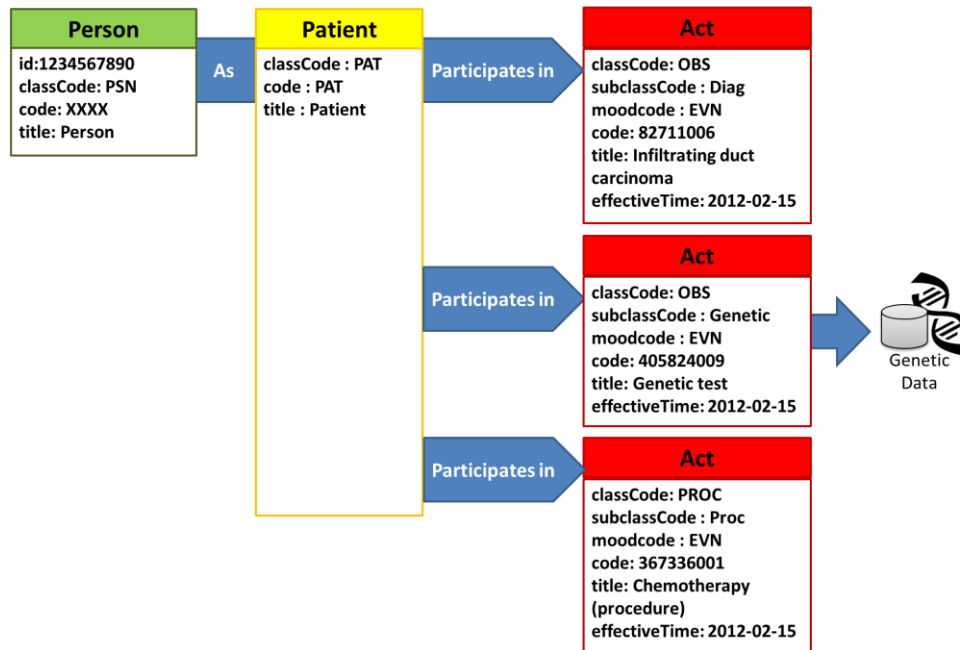
The Data Push service stores information from different data sources into the CIM representation. This service uses the normalization method of the Core Dataset service to provide a homogeneous data representation. By using the Data Push service, the CIM can

represent clinical information in the original and normalized form to ensure that no information is lost in the process.

### 2.3. Gene expression data integration

Datasets integrated within the EURECA SIL have been collected from clinical EHR systems, LIMS and CTMS integrating data of patients from clinical trials. Gene expression data extracted from clinical samples is used to explore potential interactions between clinical and genomic data.

As recommended by HL7 clinical genomics work group, genomic data could be hardly modelled as *acts* in the RIM [20]. Additionally HL7 FHIR genomics group has defined specific FHIR genomic resources [21] using SNOMED-CT and HGNC for modelling genetic tests. The authors decided to store gene expression data maintaining its original format in the CDM, by using a combination of RIM objects accordingly to the previously defined Core Dataset. A terminology guided approach has been applied, where every genomic observation in the CDM is completely defined by its field *Act code*.



**Figure 2:** Example of the Oxford Structured Breast Cancer dataset in CDM

As shown in Figure 2, patient data were annotated with Core Dataset concepts and stored in the CDM. Concretely, gene expression data were annotated with HGNC concepts linked to a HL7 RIM observation defined by the *Genetic test* SNOMED concept.

The integration of gene expression data in the SIL was a crucial task in order to facilitate the use of diagnostic classifier tools in the EURECA project. The focus of the diagnostic classifier scenario is assisting physicians in the process of diagnosing individual patients. This assistance involves tools that are trained to define diagnostic clusters using a training set of datasets. These models are then evaluated using validation datasets to assess whether new patients are clustered in the correct diagnostic category. The diagnostic classifier required input datasets that are heterogeneous for the clinical characteristics and the gene expression patterns; the algorithm is unable to distinguish diagnostic subgroups of patients if there simply is no variety in patient characteristics, for example if there is only one specific subtype of cancer present in the data. To demonstrate the suitability of the proposed approach, such heterogeneous datasets, including clinical and genomic data, were retrieved using the standard-based SIL to bind clinical and genetic information. In the Results Section the complete workflow is described, from harmonized clinical and genomic data retrieved from the SIL, to the results of the diagnostic classifier tool.

### 3. Results

Several datasets that had been shared under the EURECA project were stored and homogenized using the proposed SIL during the EURECA project. Especially relevant for genetic data integration was the Oxford dataset within the project, that is, a retrospective data collection from 219 patients from the Oxford Structured Breast Cancer dataset. This dataset is composed of clinical data of patients collected from EHR systems and gene expression data (log<sub>2</sub> measurement of mRNA abundance) of 16,814 human genes for every patient, measured using Illumina microarrays. These measurements were extracted from biopsies of breast tumor tissues. Tumor specific gene expression is highly variable within the same tumor and very different from the expression of the patient's normal tissue.

Genetic test data represents 6% of observational data and every genetic observation of a patient in the CDM contained 16K gene measurements. Although genetic test data represent a small percentage of observations (following HL7 representation depicted in Figure 2), it contains a large quantity of information. Appendix A describes the concepts used from core dataset terminologies to represent data sources within the CDM containing the Oxford dataset.

We used 219 patients from Oxford with clinical and gene expression data for the diagnostic classifier use case. A consensus clustering method (ConsensusClusterPlus) was used to group the patients based on the expression of all genes [40]. In this method, visualising the Consensus Cumulative Distribution Function (CDF) for a range of number of clusters can be used to

determine the optimal number of clusters. Although the clustering process is unsupervised, which means that the outcome value was not taken into account while clustering, distant recurrence free survival (DRFS) after 10 years of follow-up was used to test if the found clusters were prognostic, i.e., have an association with patient outcome. In Table A.1 clinical and genomic variables present in the dataset are described.

To process data coming from the SIL into R scripts, data was retrieved using SPARQL queries and then transformed it into raw dataset (in csv format). A procedure has been developed to process and analyse data for the diagnostic classifier scenario. This procedure is divided in five steps: (i) data acquisition from SIL, (ii) dataset translation, (iii) pre-processing, (iv) analysis execution, and, (v) output display.

- (i) **Data acquisition from SIL.** As every diagnosis or procedure of a patient is stored as one HL7 act – implying often several RIM objects – inside the CDM, and coded using concepts from CD, each variable required for the diagnostic classifier tool is obtained using one query built by Query Builder Service and the CIM Access Service from the SIL. Thanks to the abstraction provided by the SIL Query Builder Service, we can obtain a functional SPARQL query by asking for one (SNOMED-CT) concept. For example if we utilize ‘Node category finding (385382003)’ concept, the following SPARQL is returned from SIL Query Builder Service:

```
SELECT DISTINCT ?id ?code ?title ?patientId ?birthTime ?effectiveTime
WHERE {
    ?instPerson                hl7rim:person_id                ?patientId;
    hl7rim:person_code          '337915000';
    ?birthTime;
    hl7rim:person_participation ?instPart2.
    ?instPart2                  hl7rim:participation_act ?instAct.
    ?instAct                    hl7rim:act_code    ?code;
    hl7rim:act_title             ?title;
    hl7rim:act_id                ?id.
OPTIONAL{
    ?instAct                    hl7rim:act_effectiveTime    ?effectiveTime}
    FILTER (?code IN (isAnySubclassOf(385382003)))
}
```

This query allows retrieval of all HL7 acts stored in the CDM that are coded using the code 385382003 or any of its subsumptions, which are concrete values for the N status

finding: N0, N1, N2 and N3 Category concepts. Afterwards, when the SPARQL query is executed through the CIM Access Service, it retrieves results in XML format containing all observations of N status categories for all patients in the CD. The results obtained contain all attributes present in SPARQL query, and each row is an occurrence of an 'N status' measurement for a given patient as shown in Table 1.

**Table 1.** Example resultset for Node category finding SPARQL query

act_id	Code	Title	patientId	birthTime	effectiveTime
007c36f9-...	53623008	N1 category	63paseqzdwgq	1936-10-30	1990-12-08
			...	T00:00:00.0	T00:00:00.0
024bcfa6-...	62455006	N0 category	55af4sm4yuwt...	1941-07-03	1992-07-03
				T00:00:00.0	T00:00:00.0

Similar to clinical data observations (like 'Node category finding'), gene expression data was modelled using HL7 recommendations. For this purpose the SIL Query Builder Service is invoked using the code 'Genetic test (405824009)', obtaining the SPARQL query:

```
SELECT DISTINCT ?id ?code ?title ?patientId ?birthTime
               ?effectiveTime ?value
WHERE {
    ?instPerson      hl7rim:person_id           ?patientId;
    hl7rim:person_code      '337915000';
    hl7rim:person_birthTime ?birthTime;
    hl7rim:person_participation ?instPart2.
    ?instPart2      hl7rim:participation_act     ?instAct.
    ?instAct        hl7rim:act_code             ?code;
    hl7rim:act_title ?title;
    hl7rim:act_id    ?id.
    OPTIONAL{
        ?instAct      hl7rim:act_effectiveTime   ?effectiveTime}
        ?instAct      hl7rim:act_actObservationValues
                                ?instValues.
        ?instValues    hl7rim:actObservationValues_value
                                ?value.
    FILTER (?code      IN      (isAnySubclassOf(405824009)))
}
```

The Query Builder service returns this SPARQL query that is capable of retrieving gene expressions data stored in the SIL. Also in this case, presented in Table 2, each row of the result set corresponds to the gene expressions data of one patient.

**Table 2.** Example result set for Node category finding SPARQL query

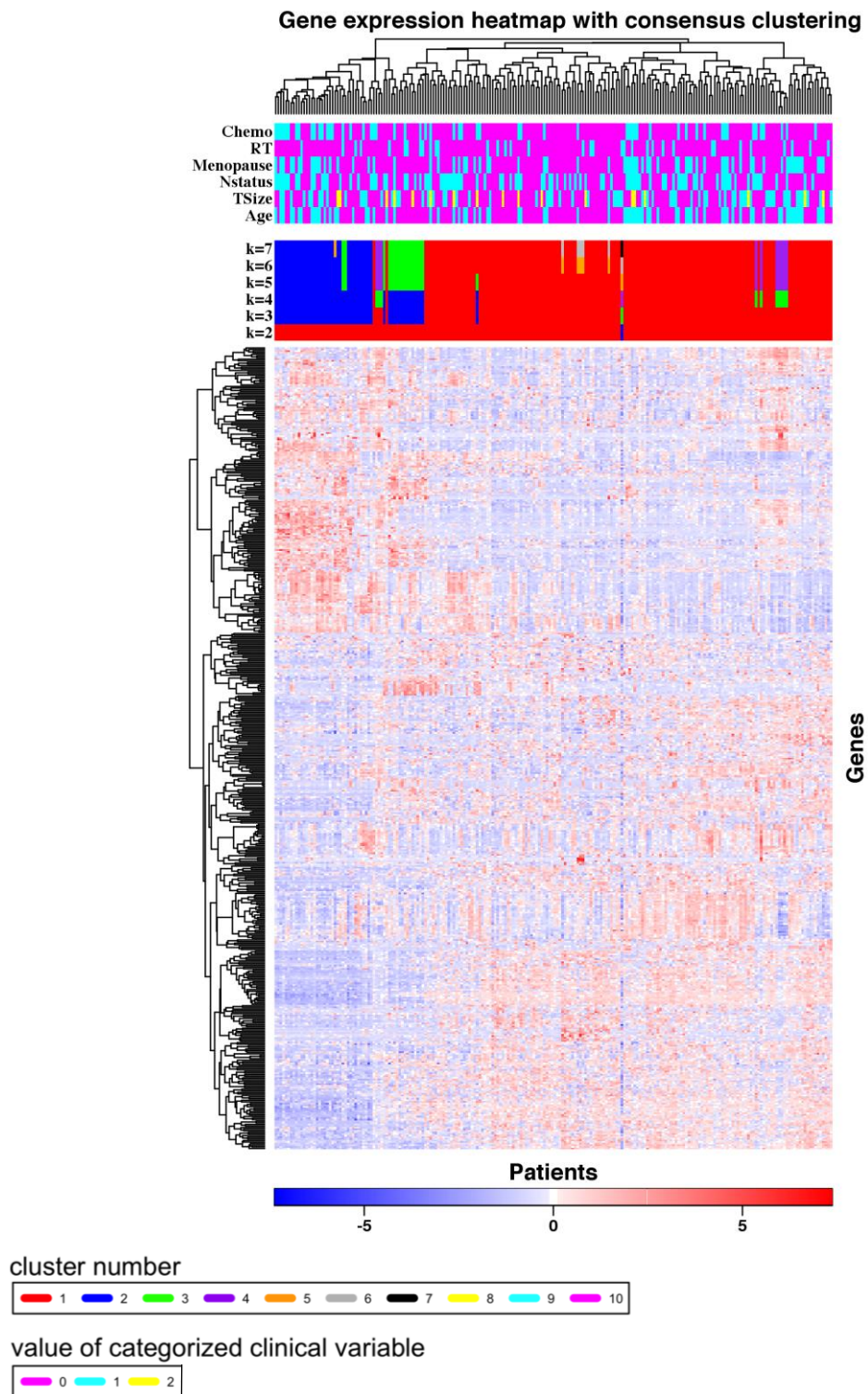
act_id	Code	title	patientId	birthTime	effective Time	Value
007c36f9-...	405824009	Genetic test	63paseqzdwgq...	1936-10-30 T00:00:00.0	1990-12-08 T00:00:00.0	ENSG0000 0091482, ENSG0000 0144834, ENSG0000 0187522, ENSG0000 0185222, ENSG0000 0008324...
007c36f9-...	405824009	Genetic test	63paseqzdwgq...	1936-10-30 T00:00:00.0	1990-12-08 T00:00:00.0	ENSG0000 0144834, ENSG0000 0187522, ENSG0000 0008324, ENSG0000 0128510, ENSG0000 0169241,....

Then, for each input variable in diagnostic classifier one different query has to be executed. Afterwards, different results obtained for each variable are combined, building a standardized cohort containing all relevant columns from each SPARQL query.

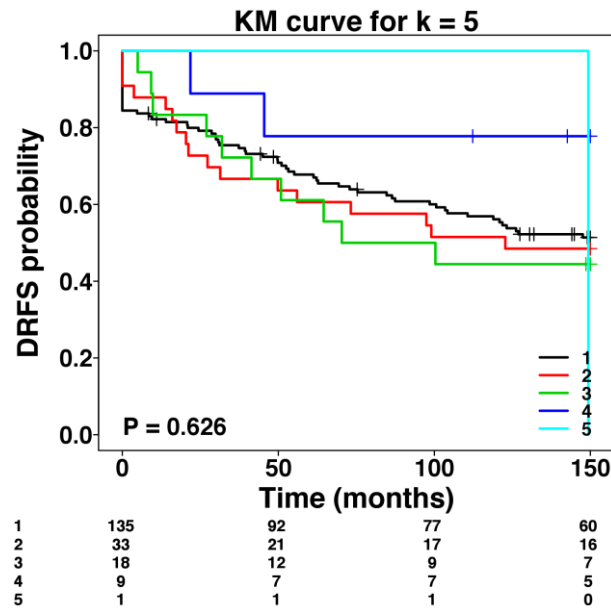
- (ii) **Dataset translation:** An intermediate step is performed to produce a dataset accepted by R scripts. SNOMED-CT codes are converted to R readable names and numerical values for clinical data. In this step gene expressions data is transformed by splitting it in different columns of the dataset.

- (iii) **Pre-processing:** The top X genes (default X = 500) with highest standard deviation for gene expression are selected, patients without outcome are excluded, missing data is inputted using expectation maximization imputation and clinical variables are categorized if necessary.
- (iv) **Analysis execution:** A consensus clustering is applied using the R-package *ConsensusClusterPlus* (v1.24.0).
- (v) **Output display:** Optimization of the number of clusters parameters and the clustering heatmap showing clustered gene expression values, assigned patient clusters and the prognostic values of these clusters.

The most important output from diagnostic classifier is the clustering heatmap, which is depicted in Figure 3. This heatmap shows the mRNA expression (normalised by z-score) of every gene for all the patients. Both axes of the figure are clustered using a hierarchical clustering, resulting in clustering of similar patients in genes on a gene expression level. Above the heatmap it can be observed which patients belong to which clusters based on the consensus clustering described above, for a range of number of clusters (in this case k=2 to 7). On top of the figure the status of the categorized clinical variables is depicted, to identify any correlation between gene expression clusters and clinical data. In the selected dataset, we can clearly see 3 clusters (red, blue and green) and the possibility to identify outlier patients that consistently form a cluster on their own. If we take the case of five clusters, which was sufficient according to the CDF visualisation, we can check whether the distant recurrence free survival is also different in these clusters (Figure 4). Kaplan-Meier curves are plotted for distant-recurrence free survival for each of the clusters from consensus clustering with the selected number of clusters.



**Figure 3:** Heatmap of gene expression (z-score), with hierarchical clustering in both directions. Clustered patients are indicated based on consensus clustering for  $k = 2$  to  $7$  with the provided colour legend and the categorical value of the clinical variables age, tumor size (TSize), nodal status (Nstatus), menopause status, radiotherapy given (RT) and chemotherapy given (Chemo) are provided at the top.



**Figure 4:** Kaplan-Meier curves for distant-recurrence free survival (DRFS) for each of the clusters from consensus clustering with selected number of clusters is 5. At the bottom is depicted how many patients are still event-free at the indicated time points. The survival curves are not significantly distinct as is tested using logrank test.

#### 4. Conclusions

This article presented a Standard-based SIL approach to integrate and facilitate the analysis of clinical and genomic data interaction in breast cancer patients. The proposed SIL allows homogeneous representation and access of patient data seamlessly facilitating the development of generic tools. This approach has been evaluated by experts within pilots and workshops during the EURECA and CIMED projects, and by European Commission experts in live demonstrations during project reviews. The proposed approach was successfully extended to other domains during the EURECA project: lung cancer, colorectal cancer, febrile neutropenia [28].

In this work, the SIL has been applied to store comprehensive genetic tests within the same structure that clinical related information is stored. The proposed integration process has been evaluated with real data from a breast cancer dataset from the Oncology Department of the University of Oxford. In this dataset, we achieved storage of information recorded from 219 patients during a clinical trial. The main contribution was to integrate the complex results of genetic tests performed on patients in the study, facilitating the access for bioinformatics tools that require genomic information for their execution.

The SIL was successfully tested for running the diagnostic classifier tool over the dataset stored in the CDM. The proposed method solves effectively most heterogeneities and integration challenges of current post-genomic clinical trial scenarios. Results show the semantic capabilities of the proposed approach, exploiting knowledge inferred from the different biomedical terminologies in the core dataset.

## Acknowledgments

This work is supported by EURECA project funded by the European Commission [grant number FP7-ICT-2011-5.3- 288048]; and the “Collaborative Project on Medical Informatics Informatics (CIMED)” funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013-2016 and the European Regional Development Funds (FEDER) [grant numbers PI13/02020, PI13/00280].

## Author Contributions:

Raul Alonso-Calvo and Sergio Paraiso-Medina wrote the paper; Raul Alonso-Calvo, Sergio Paraiso-Medina, and David Perez-Rey conceived and designed the experiments; Sergio Paraiso-Medina and Raul Alonso-Calvo performed the experiments; Sheng Yu, Ruud Van Stiphout, Marian Taylor and Francesca Buffa provided the data and contributed in diagnostic classifier experiment and analysis tools; Sergio Paraiso-Medina, Enrique Alonso-Oset and Carlos Fernandez-Lozano analyzed the data; Victor Maojo and Alejandro Pazos reviewed and validated the methodology proposed.

## References

1. McShane, L. M., Cavenagh, M. M., Lively, T. G., Eberhard, D. A., Bigbee, W. L., Williams, P. M., ... & Conley, B. A. (2013). Criteria for the use of omics-based predictors in clinical trials. *Nature*, 502(7471), 317-320.
2. Lacy, A. M., Delgado, S., Castells, A., Prins, H. A., Arroyo, V., Ibarzabal, A., & Pique, J. M. (2008). The long-term results of a randomized clinical trial of laparoscopy-assisted versus open surgery for colon cancer. *Annals of surgery*, 248(1), 1-7.
3. Ross, J. S., Lehman, R., & Gross, C. P. (2012). The importance of clinical trial data sharing: toward more open science. *Circulation. Cardiovascular Quality and Outcomes*, 5(2), 238-40. <http://doi.org/10.1161/CIRCOUTCOMES.112.965798>
4. Szalma, S., Koka, V., Khasanova, T., & Perakslis, E. D. (2010). Effective knowledge management in translational medicine. *Journal of Translational Medicine*, 8, 68. <http://doi.org/10.1186/1479-5876-8-68>
5. Richesson, R. L., & Krischer, J. (2007). Data standards in clinical research: gaps, overlaps, challenges and future directions. *Journal of the American Medical Informatics Association*, 14(6), 687-696
6. Bodenreider, O. (2008). Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of medical informatics*, 67.
7. Stang, P. E., Ryan, P. B., Racoosin, J. A., Overhage, J. M., Hartzema, A. G., Reich, C., ... & Woodcock, J. (2010). Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Annals of internal medicine*, 153(9), 600-606.
8. Murphy, S. N., Weber, G., Mendis, M., Gainer, V., Chueh, H. C., Churchill, S., & Kohane, I. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2), 124-130.
9. Beeler, G. W. (1998). HL7 Version 3—An object-oriented methodology for collaborative standards development. *International journal of medical informatics*, 48(1), 151-161.
10. Bender, D., & Sartipi, K. (2013, June). HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. In *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on* (pp. 326-331). IEEE.
11. Bernardini, A., Alonzi, M., Campioni, P., Vecchioli, A., & Marano, P. (2002). IHE: integrating the healthcare enterprise, towards complete integration of healthcare information systems. *Rays*, 28(1), 83-93.
12. Fleurence, R. L., Curtis, L. H., Califf, R. M., Platt, R., Selby, J. V., & Brown, J. S. (2014). Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association*, 21(4), 578-582.
13. Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121, 279.

14. McDonald, C. J., Huff, S. M., Suico, J. G., Hill, G., Leavelle, D., Aller, R., ... & Williams, W. (2003). LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry*, 49(4), 624-633.
15. Stewart, B. and Wild, C.P. (eds.), International Agency for Research on Cancer, WHO. (2014) World Cancer Report 2014 [Online]. Available from:
16. Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., & Wain, H. (2001). The HUGO gene nomenclature committee (HGNC). *Human genetics*, 109(6), 678-680.
17. Canuel, V., Rance, B., Avillach, P., Degoulet, P., & Burgun, A. (2014). Translational research platforms integrating clinical and omics data: a review of publicly available solutions. *Briefings in bioinformatics*, bbu006.123213
18. Warner, J. L., Jain, S. K., & Levy, M. A. (2016). Integrating cancer genomic data into electronic health records. *Genome Medicine*, 8, 113. <http://doi.org/10.1186/s13073-016-0371-3>
19. Athey, B. D., Braxenthaler, M., Haas, M., & Guo, Y. (2013). tranSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. *AMIA Summits on Translational Science Proceedings*, 2013, 6–8.
20. Amnon Shabo, HL7 Clinical Genomics Work Group - May 2016. Available at url: <http://www.hl7.org/special/committees/clingenomics/>
21. Alterovitz G et al, SMART on FHIR Genomics: facilitating standardized clinico-genomic apps. *J Am Med Inform Assoc*. 2015 Nov;22(6):1173-8. doi: 10.1093/jamia/ocv045
22. CDISC Clinical Data Interchange Standards Consortium, <http://www.cdisc.org>.
23. CDISC Study Data Tabulation Model (SDTM); <https://www.cdisc.org/standards/foundational/sdtm>
24. CDISC Pharmacogenomics /Genetics PGx; <https://www.cdisc.org/standards/foundational/pgx>
25. McConnell, P., Dash, R. C., Chilukuri, R., Pietrobon, R., Johnson, K., Annechiarico, R., & Cuticchia, A. J. (2008). The cancer translational research informatics platform. *BMC Medical Informatics and Decision Making*, 8, 60. <http://doi.org/10.1186/1472-6947-8-60>
26. Spjuth O, Krestyaninova M, Hastings J, et al. Harmonising and linking biomedical and clinical data across disparate data archives to enable integrative cross-biobank research. *European Journal of Human Genetics*. July 2016; 24(4):521-528. doi:10.1038/ejhg.2015.165
27. Kho AN, Rasmussen LV, Connolly JJ, Peissig PL, Starren J, Hakonarson H, et al. Practical challenges in integrating genomic data into the electronic health record. *Genet Med*. 2013;15:772–8.
28. EURECA Project: URL: <https://eurecaproject.eu>
29. Ibrahim, A., Bucur, A., Dekker, A., Marshall, M. S., Perez-Rey, D., Alonso-Calvo, R., & Mehta, K. (2014, November). Analysis of the Suitability of Existing Medical Ontologies for Building a Scalable Semantic Interoperability Solution Supporting Multi-site Collaboration in Oncology. In *Bioinformatics and Bioengineering (BIBE)*, 2014 IEEE International Conference on (pp. 204-211). IEEE.
30. Broekstra, J., Kampman, A., & Van Harmelen, F. (2003). Sesame: An architecture for storing and querying RDF data and schema information. *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*, 197.
31. Fritz, A., Percy, C., Jack, A., Shanmugaratnam, K., Sobin, L., Parkin, D. M., & Whelan, S. (2000). International classification of diseases for oncology (No. Ed. 3). World Health Organization.
32. Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*, 32(suppl 1), D258-D261.
33. Sioutos, N., de Coronado, S., Haber, M. W., Hartel, F. W., Shaiu, W. L., & Wright, L. W. (2007). NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of biomedical informatics*, 40(1), 30-43.
34. Moratilla, J. M., Alonso-Calvo, R., Molina-Vaquero, G., Paraiso-Medina, S., Perez-Rey, D., & Maojo, V. (2012). A data model based on semantically enhanced HL7 RIM for sharing patient data of breast cancer clinical trials. *Studies in health technology and informatics*, 192, 971-971.
35. Priyatna, Freddy, Raúl Alonso-Calvo, Sergio Paraiso-Medina, Gueton Padron-Sanchez and Óscar Corcho. "R2RML-based Access and Querying to Relational Clinical Data with Morph-RDB." *SWAT4LS* (2015).

36. Benson, T. "Principles of Health Interoperability HL7 and SNOMED". London: Springer-Verlag; 2010.
37. Spackman KA. Normal forms for description logic expressions of clinical concepts in SNOMED RT. Proceedings of the AMIA Symposium. 2001:627-631.
38. Cheetham E, H. Dolin R, Markwell D, Curry J, Gabriel D, Hausam R, Knight B, Rector A, Spackman K, Townend I. "Using SNOMED CT in HL7 v3 Implementation Guide", Release 1.5. 2008.
39. Paraiso-Medina, S., Perez-Rey, D., Bucur, A., Claerhout, B., & Alonso-Calvo, R. (2015). Semantic normalization and query abstraction based on SNOMED-CT and HL7: supporting multicentric clinical Trials. Biomedical and Health Informatics, IEEE Journal of, 19(3), 1061-1067.
40. Wilkerson MD, Hayes DN (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics, 26(12), 1572-3.

## Appendix A. Data characterization of the clinico-genomic data source integrated

**Table A.1.** Core Dataset concepts present in Oxford breast cancer dataset within CDM

CD Concept Code	CD Label	% Patients incl. concept	% of total data
443527007	Number of lymph nodes involved by malignant neoplasm (observable entity)	100%	6,10%
444025001	Number of lymph nodes examined (observable entity)	100%	6,10%
263605001	Tumor size (observable entity)	100%	6,10%
405824009	Genetic test	98,63%	6,02%
106221001	Genetic finding (finding)	95,89%	5,85%
108290001	Radiation oncology AND/OR radiotherapy (procedure)	84,02%	5,12%
178294003	Axillary lymph nodes sampling (procedure)	77,17%	4,71%
82711006	Infiltrating duct carcinoma (morphologic abnormality)	74,43%	4,54%
289903006	Menopause present (finding)	67,12%	4,09%
64368001	Partial mastectomy (procedure)	66,67%	4,07%
309542002	Endocrine therapy (procedure)	58,45%	3,56%
62455006	N0 category (finding)	57,99%	3,54%
161917009	Recurrence of problem (finding)	56,16%	3,43%
399879007	Malignant epithelial neoplasm - category (morphologic abnormality)	56,16%	3,43%
399350006	Under follow-up (finding)	51,14%	3,12%
30893008	M0 category (finding)	51,14%	3,12%
14799000	Neoplasm, metastatic (morphologic abnormality)	48,86%	2,98%
445150007	Surviving free of recurrence of neoplastic disease (finding)	43,84%	2,67%
53623008	N1 category (finding)	42,01%	2,56%
1663004	G2 grade (finding)	41,55%	2,53%
289904000	Menopause absent (finding)	32,88%	2,01%
61026006	G3 grade (finding)	31,05%	1,89%
367336001	Chemotherapy (procedure)	25,11%	1,53%

172043006	Simple mastectomy (procedure)	23,74%	1,45%
54102005	G1 grade (finding)	19,18%	1,17%
89740008	Lobular carcinoma (morphologic abnormality)	12,33%	0,75%
444057000	Infiltrating carcinoma with ductal and lobular features (morphologic abnormality)	9,13%	0,56%
122548005	Biopsy of breast (procedure)	7,31%	0,45%
234254000	Excision of axillary lymph nodes group (procedure)	1,83%	0,11%
392021009	Lumpectomy of breast (procedure)	1,83%	0,11%
4631006	Tubular adenocarcinoma (morphologic abnormality)	1,83%	0,11%
32913002	Medullary carcinoma (morphologic abnormality)	1,37%	0,08%
72495009	Mucinous adenocarcinoma (morphologic abnormality)	0,91%	0,06%

---