Executive Summary for deliverable D4.1: Inventory of information and knowledge models and Definition of EHR4CR Information Models

Project acronym: EHR4CR
Project full title: Electronic Health Records for Clinical Research
Grant agreement no.: 115189
Budget: 16 million EURO
Start: 01.03.2011 - End: 28.02.2015
Website: www.ehr4cr.eu

Coordinator: AstraZeneca
Managing Entity: EURORec
Document description

<table>
<thead>
<tr>
<th>Deliverable no:</th>
<th>D4.1 (executive summary)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deliverable title:</td>
<td>Inventory of information and knowledge models and Definition of EHR4CR Information Models</td>
</tr>
<tr>
<td>Description:</td>
<td>First deliverable of Work Package 4</td>
</tr>
<tr>
<td>Status:</td>
<td>Final</td>
</tr>
<tr>
<td>Version:</td>
<td>0.2</td>
</tr>
<tr>
<td>Date:</td>
<td>14/08/2012</td>
</tr>
</tbody>
</table>

Document history

<table>
<thead>
<tr>
<th>Date</th>
<th>Revision</th>
<th>Author(s)</th>
<th>Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>13/08/2012</td>
<td>0.1</td>
<td>C. Daniel, S. Hussain</td>
<td>Initial draft</td>
</tr>
<tr>
<td>14/08/2012</td>
<td>0.2</td>
<td>C. Daniel, S. Hussain</td>
<td>Final version</td>
</tr>
</tbody>
</table>

Table of Content

1. Introduction .................................................................................................................. 4
   1.1. Objective and context ............................................................................................ 4
   1.2. Overview ............................................................................................................... 4

2. Task 4.1: Inventory of information and knowledge models ........................................ 5
   2.1. Task 4.1 (a): Inventory of standard-based information and knowledge models -- State of the Art 5
       2.1.1 Method............................................................................................................... 6
       2.1.2 Results............................................................................................................. 6
       2.1.3 Conclusion of the state of the art .................................................................. 8
   2.2. Task 4.1 (b): Pharma survey (clinical trials & eligibility criteria) ......................... 8
       2.2.1 Method............................................................................................................... 8
       2.2.2 Results............................................................................................................. 9
       2.2.3 Conclusion of the pharma survey and formalization of eligibility criteria ............ 10
   2.3. Task 4.1 (b): Hospital survey (local EHRs/CDWs information models & terminologies) ...... 11
       2.3.1 Method............................................................................................................... 11
       2.3.2 Results............................................................................................................. 11
       2.3.3 Conclusion of the hospital survey .................................................................... 12

3. Task 4.2: Definition of EHR4CR information models ..................................................... 13
   3.1. Methods.................................................................................................................... 13
   3.2. Result...................................................................................................................... 13
3.2.1 Clinical Information model ............................................................................................................. 13
3.2.1.1 EHR4CR clinical information model (v2.0) ............................................................................. 14
3.2.1.2 EHR4CR metadata repository ............................................................................................... 14
3.2.1.3 EHR4CR terminology .......................................................................................................... 15
3.2.2 Query model .................................................................................................................................. 15
3.3 Conclusion of the EHR4CR semantic interoperability framework .................................................. 15
4 References ......................................................................................................................................... 17
1 Introduction

A major barrier to repurposing routinely collected clinical data for clinical research is that real-world information systems in both domains – patient care and clinical research – use different information models and terminology systems. The EHR4CR (Electronic Health Records for Clinical Research) platform will implement four use cases – protocol feasibility testing, patient identification and recruitment for clinical trials, supporting clinical trial execution and adverse event reporting – to be demonstrated by 10 pilots in 5 European countries.

1.1. Objective and context

WP4 (Semantic interoperability) aims to design and implement a standard-based, expressive and scalable semantic interoperability framework allowing dynamic mappings between the data structures and semantics of the two data usage contexts: patient care on one side and clinical research on the other side. WP 3 (WP 3: Architecture and Integration) defines how the tools and services developed by WP 4 (Semantic interoperability) will be integrated with Data Protection, Privacy & Security services (WP5) and end-user Platform Services (WP6) into a loosely coupled service platform.

The core of the EHR4CR semantic interoperability framework (or “EHR4CR pivot representation”) is a shared conceptual reference model (EHR4CR information model) acting as a global as view model to correlate the schemas and concepts from varying sources managed and implemented through the use of the EHR4CR meta data repository. A shared terminology (EHR4CR terminology) provides the codes of the encoded properties of the classes of the EHR4CR information model.

EHR4CR is not just seeking the ideal pivot representation, but the optimal balance between perfection and an affordable and scalable integration of heterogeneous EHR systems to the EHR4CR platform.

This deliverable D4.1 (M12) “Inventory of information and knowledge models. Definition of EHR4CR information models” is the first step in coming to that optimal understanding. It describes the result of the task 4.1 (Inventory of information and knowledge models and Task 4.2 (Definition of EHR4CR Information Models).

The current focus is patient eligibility determination (protocol feasibility testing as specified in EHR4CR_Protocol_Feasibility_SRS_v1.0 and patient recruitment). Figure 1 provides an overview of semantic interoperability issues for patient eligibility determination.

1.2. Overview

Part 1 (Task 4.1) provides an inventory of standard-based information and knowledge models. Task 4.1a consisted of an inventory of state of the art of information models and terminology standards in use within healthcare and clinical research and of patient care/clinical research integration initiatives. Task 4.1b consisted of investigating the data structures and terminologies used in pilot sites - pharma companies and university hospitals - in order to precisely define the scope that needs to be addressed by the EHR4CR information model and terminology. Although not covering the complete marketplace, these pilot sites and companies present a diverse enough spectrum of de
Part 2 (Task 4.2) describes the current version of the **EHR4CR semantic interoperability framework**.

![EHR4CR communication models](image)

**Figure 1.** Overview of semantic interoperability issues for patient eligibility determination (protocol feasibility and patient recruitment)

2. **Task 4.1: Inventory of information and knowledge models**

2.1. **Task 4.1 (a): Inventory of standard-based information and knowledge models -- State of the Art**

The objective of task 4.1a was to review the **information models and terminology standards** in use within healthcare and clinical research and existing **patient care/clinical research integration initiatives**.
2.1.1 Method

The scope of the state of the art addresses: i) Healthcare terminologies/ontologies & information models used in both patient care & clinical research areas and ii) Patient care/clinical research integration initiatives.

We reviewed the main standard Healthcare Information Technology Standards Development Organizations (SDOs) (HL7, DICOM, IHE, CEN TC 251, ISO TC215, w3C, IHTSDO, etc) websites to identify biomedical terminologies and ontologies, healthcare information models (e.g CDA templates, 13606 archetypes, etc), core data sets (e.g CDISC SHARE) that could be relevant for the EHR4CR project. We reviewed HL7, CDISC, IHE & e-clinical forum websites in order to identify initiatives aiming at using clinical information systems (EHRs&CDWs especially in hospitals) in order to support clinical research. We searched Medline for articles dedicated to semantic interoperability for patient care and research (clinical research or epidemiology) integration.

2.1.2 Results

- Healthcare terminologies/ontologies & information models
  In the domain of patient care, several large-scale efforts have been underway for over a decade with the goal of specifying both the structure and the semantics of patient clinical information in a manner that enables computable semantic interoperability between diverse systems. ISO EN 13606, the openEHR Foundation and HL7 RIM and Clinical Document Architecture are the three major contributions to the interoperability of clinical information.

  In the domain of clinical research, the CDISC organization has developed a number of platform-independent standards that support the electronic acquisition, exchange, regulatory submission, and subsequent archiving of clinical research data. In particular, in 2001, CDISC published the first version of its Operational Data Model (ODM), a specification that defined the organization, structure and syntax of data captured for analysis and reporting over the course of a clinical trial [CDISC12a]. Recently, the Clinical Data Acquisition Standards Harmonization (CDASH) initiative specified the unambiguous semantics of a number of common data elements that are deemed “common” to all trials. As such, CDASH represents a significant first-step in achieving cross-trial semantic interoperability.

  Recent efforts have focused on bringing the various international efforts done for patient care and clinical research into closer alignment. In 2004, CDISC and HL7 – along with the National Cancer Institute and the FDA –produced the Biomedical Research Integrated Domain Group (BRIDG) model which, on one side, contains representations of clinical research data with underlying mappings to the HL7 RIM and, on the other side, covers a superset of the scope defined by CDASH [Fridsma08]. CDISC has committed to migrating all of its standards to be expressed using BRIDG semantics, and the HL7 Regulated Clinical Research Information Management (RCRIM) Work Group within HL7 is committed to developing all of its message specifications in the context of BRIDG compliance.

  - Patient care/clinical research integration initiatives

---

1 SOA group: C.Daniel, D.Kalra, D.Ouagne, D Schwarz-hertzner, J.James, K.Forsberg, M.Cuggia, M.McGilchrist, R.Bache, R.Kush, S.Mate
Recently, a number of investigators have examined the various roles that EHR systems might assume in the clinical research context [Turisco05, Powel05, West09, Ohmann09, Prokosch09; Breil09]. Some [Ohmann07, Dugas10, Weng10] have pointed out that EHR data may be useful during trial design by providing trial planners with a better understanding of the available cohorts based on the trial’s Inclusion and Exclusion criteria and developed Clinical Trial Recruitment Support Systems (CTRSS) (EHR4CR use cases 1 and 2). Others [Williams03, Kush07, Murphy07, El Fadly07, El Fadly11] have specifically discussed the lessening of the burden and optimization clinical trial data collection through the targeted re-purposing of EHR data during a trial's execution phase and adverse event detection and reporting (EHR4CR use case 3 and 4).

- Semantic interoperability framework for patient eligibility determination (use cases 1 & 2)

The use of EHRs or CDWs for eligibility determination requires the definition of a formal representation of (usually free-text) eligibility rules and a semantic matching solution between clinical constraints of expressed in the eligibility criteria and patient data routinely collected in heterogeneous clinical systems. Weng and coll. recently surveyed the literature about computable representations of eligibility criteria and defined a conceptual framework that can serve as an evaluation matrix for future users or developers of computable eligibility criteria to select relevant standards in this area. Developing a system for eligibility determination requires the definition of a formal representation of eligibility criteria based on:

i) A query language representing executable eligibility rules

Languages of varying expressiveness have been used to represent the logic of eligibility criteria. The development of ad hoc expression languages (EON [Musen 96], SAGE [Tu07], and ERGO [Sim04]) is driven by use cases instead of any theoretical basis. In contrast, the other languages have a theoretical basis and formal foundation (Arden Syntax; variants of logic-based languages, such as Structured Query Language (SQL), description logic (DL), object-oriented query and expression languages, such as Object Constraint Language (OCL), GELLO [GELLO12a, GELLO12b, Sordo04, Mei11], ruleML and SBVR.

ii) A patient information model

Systems developed after the year 2000 (including GUIDE, GLIF3, SAGE, ERGO, CRFQ, Patel’s and Lonsdale’s systems) largely adopted some sophisticated models, providing an abstraction layer for EHRs or CDWs called Virtual Medical Record (VMR). Some of these models are based on the HL7 Reference Information Model (RIM), with varying degrees of adoption (including for instance, only one Observations class) [Johnson01, Jenders97, Lonsdale07]. Other initiatives rely on CEN 13606 interoperability standards [Nies07, Dziuballe11]. A current trend in the design of a shared patient information model is to propose a generic object-oriented data model (UML model) associated with data structures that are already defined and standardized. Recent efforts in both patient care and clinical research consist in defining metadata and vocabulary standards for clinical information and thereby in building Common Data Elements (CDEs)(also called metadata repositories or item banks) [Nadkami06]. The CDEs are structured data elements, consisting of precisely defined questions and answers (e.g. NCI’s caDSR [Covitz03, Warzel03], caMatch approach in the ASPIRE project, CDISC SHARE [CDISC12b]).

iii) An appropriate clinical terminology to facilitate mapping from eligibility concepts to patient data.
Achieving semantic interoperability requires the use of both standard data structures (information models and data elements) and common concepts and their interpretation [Oemig10]. A range of clinical terminologies are needed to collectively represent the variety of clinical statements. Multiple clinical terminologies will be needed to support representation for different data sources, including diagnosis, findings, familial and medical history, lab tests, medication, etc.

- Semantic interoperability for data collection (use cases 3 & 4)
Two IHE profiles dedicated to research and public health – as proposed by both the IHE Quality Research and Public Health (QRPH) and Information Technology Infrastructure (ITT) domains – address the issue of multi-vendor, scalable interoperability required for multicenter trials [IHE ITI RFD12, IHE QRPH CRD12]. The Retrieve Form for Data-capture (RFD) integration profile [IHE ITI RFD12] combined with the Clinical Research Document (CRD) content profile [IHE QRPH CRD12] collectively provide a conceptual framework for implementing the “single-point-of-data-collection” approach to EHR/CDMS integration.

2.1.3 Conclusion of the state of the art
The focus of the EHR4CR project will be first on eligibility determination (use cases 1 & 2). Supporting protocol feasibility studies and/or eligibility determination requires the definition of computable representation of both eligibility criteria and routinely collected clinical data in order to support automated retrieval of patient numbers and/or of individuals who are eligible for a given clinical trial.

We clearly stated that (semi-)automated eligibility determination will require the definition of a query language representing executable eligibility rules (query model), a common clinical information model (EHR4CR clinical information model) and a common clinical terminology (EHR4CR clinical terminology).

The EHR4CR semantic interoperability framework (EHR4CR information models and terminology services) shall be developed consistently with key standard-based content & integration profiles developed internationally.

The EHR4CR consortium will report to the relevant standards bodies and initiatives (e.g. CDISC, HL7, DICOM, IHTSDO, ISO, CEN, IHE) if any limitations in their existing standards have been identified and extensions are needed.

2.2. Task 4.1 (b): Pharma survey (clinical trials & eligibility criteria)
Task 4.1b consisted in investigating the data structures and terminologies that could be used to represent eligibility criteria.

2.2.1 Method
WP7 collected a set of clinical trials in the domain areas of the project (cardiovascular diseases, oncology, diabetes, inflammatory diseases, neurologic diseases, respiratory diseases, etc) sponsored by the pharma companies of the EHR4CR project and running in more than one of the university hospitals.

---

2 Eligibility criteria analysis group (Pharma inventory group) Responsible : C Daniel & R.Bache – D. Acosta, M-C.Jaulent, T.Dart, J.James, M.Cuggia, J.Doods, B.Breil, D.Schwarz-hertzner, L.Toldo
Eligibility criteria were (manually) **pre-processed** using the ERGO methodology and the characterization framework of [Weng10] enabling the indexing, classification, retrieval, and usage of eligibility criteria. Attributes (or “data elements”) and their value range or value sets were manually extracted. Data types are based on ISO 21090 data types. Units refer to UCUM. The clinical concepts corresponding to the “data elements” and to the values in value sets were encoded using the **EHR4CR terminology**.

In the EHR4CR project, we will use **Object Constraint Language (OCL)** as the formal representation of eligibility criteria for execution on distributed heterogeneous data bases (EHRs/CDWs). Later on, during the project, other candidate expression languages (ERGO, SPARQL, RuleML and SBVR) will be considered in parallel experiments.

### 2.2.2 Results

A set of 43 clinical trials, running in more than one of the university hospitals, were selected by the EFPIA partners. We decided to first consider a **sub-set of 10 clinical trials** of the 43 as the material for the initial proof of concept (shown in Table 1). The set provide to **269** eligibility criteria.

Table 1: Sub-set of 10 clinical trials selected as the material for the proof of concept of the foreseen impact of the EHR4CR platform during feasibility studies (WP7 round 2 data export).

<table>
<thead>
<tr>
<th>Internal Nr/Code</th>
<th>Study</th>
<th>EFPIA Partner</th>
<th>Disease Area</th>
<th>Ap- hp</th>
<th>FAU</th>
<th>HUG</th>
<th>KCL</th>
<th>MU</th>
<th>W</th>
<th>US6</th>
<th>UCL</th>
<th>Univ</th>
<th>wan</th>
<th>UoG</th>
<th>WW</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>11899</td>
<td>Bayer</td>
<td>Cardiovascular</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>20050182</td>
<td>Amgen</td>
<td>Oncology</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>27919</td>
<td>Merck</td>
<td>Nervous system disorders</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>BIO1111482</td>
<td>GSK</td>
<td>Oncology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>CENA713B2315</td>
<td>Novartis</td>
<td>Neurology</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>COU-AA-301</td>
<td>Janssen</td>
<td>onco</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>D3191C00009</td>
<td>AstraZeneca</td>
<td>CV/Arrhythmias</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D4320C00015</td>
<td>AstraZeneca</td>
<td>Oncology</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFC11785</td>
<td>Sanofi</td>
<td>Oncology</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NC251113</td>
<td>Roche</td>
<td>Cardiovascular and Metabolic</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td></td>
<td>25</td>
</tr>
</tbody>
</table>

The EFPIA partners have manually pre-processed a sub-set of 99 free-text eligibility criteria and transformed them in combinations of elementary queries that were semantically annotated using concepts from reference clinical terminologies (such as ICD-10, LOINC, SNOMED CT, etc) available in the **EHR4CR terminology**. A subset of these elementary queries has been formally represented in OCL (see Table 2).
Table 2: An example of a free text eligibility criteria and its corresponding formal OCL query

<table>
<thead>
<tr>
<th>Initial criteria</th>
<th>OCL Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient with histologically proven stage III cutaneous melanoma presenting with</td>
<td><code>select * from Patient where histologicallyProvenStageIII_cutaneousMelanoma and macroscopicLymphNodeInvolvementSuitableForSurgery</code></td>
</tr>
<tr>
<td>macroscopic lymph node involvement suitable for surgery”</td>
<td></td>
</tr>
</tbody>
</table>

Existing tools were used to edit and validate the OCL syntax. Model-driven engineering approach was used to support the transformation of OCL statements into SQL statements including involvement of terminology services for mapping central to local codes and semantic expansion (see Figure 2).

**Formal representation of eligibility criteria using OCL**

![Diagram](https://via.placeholder.com/150)

*Figure 2: Formal representation of pre-processed eligibility criteria into OCL queries that are transformed into SQL queries operating on distributed heterogeneous EHRs/CDWs*

### 2.2.3 Conclusion of the pharma survey and formalization of eligibility criteria

In the EHR4CR project, we proposed a methodology to provide and evaluate a formal representation of eligibility criteria. Eligibility criteria are manually pre-processed and transformed in combinations of elementary queries formally represented in OCL\(^3\). Later on, during the project, other candidate expression languages (ERGO, SPARQL, RuleML and SBVR) will be considered in parallel experiments.

\(^3\) The elementary queries are also represented using an ad-hoc object-oriented languages (ECLECTIC) intended for presenting queries to human readers but not for execution.
A **sub-set of 10 clinical trials** corresponding to **269** eligibility criteria have been pre-processed by pharma companies. 99 free-text eligibility criteria have been represented in OCL.

### 2.3. Task 4.1 (b): Hospital survey (local EHRs/CDWs information models & terminologies)

Task 4.1b consisted in investigating the data structures and terminologies used in pilot sites (university hospitals).

#### 2.3.1 Method

The WP4 Hospital Survey was designed to establish the content, structure, semantics and some operational characteristics of the data sources available to EHR4CR at each hospital site.

Each source was surveyed for 9 categories of data: Demography, Diagnosis, Procedure, Laboratory, Anatomic pathology, Medication, Finding, Encounter and Organisation. For each category of data (when available) the total number of records and patient counts is requested, along with the first year the category of data was generally available. Finally, for data elements generally found within each category the availability, structure and semantics of the element are requested using two templates: one for the data element itself, and the other for one or more value sets associated with the data element.

#### 2.3.2 Results

There is one data source at each site. Three of the sites are using i2b2. Six of the sites have data sources available now, including all the i2b2 warehouses. Six of the data sources have good coverage of the disease categories of interest, while the remainder has partial coverage. KCL’s data source relates to cancer only. Four data sources have only recent coverage, while 3 data sources have more than 5 years coverage (see Table 3).

#### Table 3: General information for each site

<table>
<thead>
<tr>
<th>General information</th>
<th>Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>APHP</td>
</tr>
<tr>
<td>Number of sources at site?</td>
<td>1</td>
</tr>
<tr>
<td>Type of data source?</td>
<td>i2b2</td>
</tr>
<tr>
<td>Status?</td>
<td>AV</td>
</tr>
<tr>
<td>Covers CVD?</td>
<td>√</td>
</tr>
<tr>
<td>Covers cancer?</td>
<td>√</td>
</tr>
<tr>
<td>Covers diabetes?</td>
<td>√</td>
</tr>
<tr>
<td>Covers inflammatory</td>
<td>√</td>
</tr>
</tbody>
</table>

---

4 Hospital inventory group: C Daniel, T.Dart, M.McGilchrist, D Schwarz-hertzner
disease?

| Covers neurological disease? | √ | √ | √ | x |
| Covers respiratory disease? | √ | √ | √ | x |
| Covers other diseases? | Renal |

At the level of the 9 categories defined, the survey requested the extent of available data and its temporal coverage. The survey also requested the granularity of timestamps on the data on the assumption that this would be similar for all data elements within a category (see Table 4).

Table 4: Available data and temporal coverage by category

<table>
<thead>
<tr>
<th>Data category</th>
<th>Site</th>
<th>APHP</th>
<th>FAU</th>
<th>HUG</th>
<th>KCL</th>
<th>MUW</th>
<th>U936</th>
<th>UCL</th>
<th>UNIVDUN</th>
<th>UOG</th>
<th>UOM</th>
<th>WWU</th>
</tr>
</thead>
</table>

2.3.3 Conclusion of the hospital survey

The hospital survey provides the list of data elements, value sets and terminology used in local EHRs or CDWs of the 11 pilot sites. It shows reasonable coverage across the sites for demographics, diagnoses, procedures, encounters and laboratory text information, but reduced levels of information availability for medication, pathology and findings information.
3. **Task 4.2: Definition of EHR4CR information models**

The objective of Task 4.2 was to define the EHR4CR semantic interoperability framework for protocol feasibility studies scenario. This semantic interoperability framework includes clinical information models and a query model.

### 3.1. Methods

Defining a standard representation of clinical data for the EHR4CR platform follows a three steps methodology:

- **Step 1:** Designing the **conceptual model**
- **Step 2:** Identifying the **standard-based reference information models** covering the broad scope of the EHR4CR project. These “source” models consists in reference healthcare information models identified by the state of the art (task 4.1a) as well as real-world data models of EHRs or CDWs identified by the pilot site inventory (task 4.1b).
  - Identifying **HL7 v3 models** (in MIF format), archetypes and **CDA templates** that are relevant in the scope of EHR4CR i.e corresponding to the following identifying from T4.1b (inventory of information model in pilot sites)
  - Identifying interesting **source models of CDWs such as i2b2**.
- **Step 3:** Designing the **logical model** using a model-driven engineering approach. Deriving the **EHR4CR platform-independent information model** from standard-based reference information models (see Figure 3).
- **Step 4:** Linking the information model to Common Data Elements and to reference terminologies available in the **EHR4CR terminology**.

### 3.2. Result

#### 3.2.1 Clinical Information model

The key characteristics/requirements of the Clinical Information Models are

- Good level of content coverage & complexity
- Addresses the Protocol Feasibility Services (PFS) specifications
- Consistent with the current status of existing EHRs, CDWs
- Covers various medical fields (cardiovascular diseases, oncology, diabetes, inflammatory diseases, neurologic diseases, respiratory diseases, etc)
- Covers various categories of data (diagnoses and procedures collected as part of DRG systems, tests results (lab, anatomic pathology, radiology, etc), patient care coordination data elements (findings related to vital signs, medical history, immunization, etc.), etc).
- Extensible (possible extension for Patient Recruitment Services (PRS), Clinical Trial Data Capture Services (DCS) and for Adverse Event Detection and Reporting Services (ADR))
- Implementable and scalable, based on established health informatics and IT standards. Each model should contain the minimum properties needed to meet the requirements
- Cost of mapping CDW systems should be minimised
- Models should initially be specified and ratified as logical models (e.g. UML class diagrams)
- The learning curve and skills set requirements for the wider market place should be minimised
- Maximise use of established technologies

---

5 EHR4CR information model and CDW working group: C Daniel, M McGilchrist, J James, D Ouagne, D Schwarz-hertzner
• Allows an implementable interface to the query model to define query across the EHRs using the common clinical terminology.

3.2.1.1 EHR4CR clinical information model (v2.0)

3.2.1.1 EHR4CR clinical information model (v2.0)  

![Diagram of EHR4CR clinical information model](image)

Figure 3: EHR4CR multi-dimensional model (see Appendix for detailed description)

At the center of this model, the **ClinicalStatement** class is a generic abstract class further specialized in specific sub-classes used to represent medical facts that need to be described in the context of the design of a study.

A set of dimensions (**Subject, Encounter, Participation**) are used to represent information about the context of these medical facts.

The class ‘Subject’: this dimension represents the subject of the medical fact (e.g. the patient)

The class ‘Encounter’: This dimension represents the administrative context of the medical fact (this dimension is based on the I2B2 model)

The class ‘Participation’: This dimension represents all other dimensions invoked in the description of the medical fact.

A fifth and final class, named ‘ClinicalStatementRelationship’: This class is not a dimension but it points to an instance of the class ‘ClinicalStatement’ which can make reference to several other instances of the class ‘ClinicalStatement’ through instances of the class ‘ClinicalStatementRelationship’

3.2.1.2 EHR4CR metadata repository

Like in the caMatch approach, we shall define the **Common Data Elements (CDEs)** useful in the scope of the EHR4CR project. The CDEs are structured data elements, consisting of precisely defined questions and answers to instantiate the generic high level EHR4CR information model. Therefore, for a specific clinical statement (defined by a given “code” in the ClinicalStatement class), the corresponding data element provides explicit representation of the data type of the attribute “value” (and the corresponding information such as the range (value set, unit, etc).
The EHR4CR CDEs shall be used together with query languages to define eligibility criteria constructs that can be evaluated against EHR data. It is understood that there will be criteria that either cannot be formalized as computable expressions or will not have EHR data to support automated evaluation of CDEs. The EHR4CR MDR shall be aligned to other MDR (caDSR, openMDR, eMERGE repository, etc). These efforts will be done by the EHR4CR consortium, in collaboration with CDISC partner, as a contribution to the CDISC SHARE project.

### 3.2.1.3 EHR4CR terminology

The EHR4CR terminology shall integrate a range of clinical terminologies that are needed to collectively represent the variety of clinical statements (including diagnosis, findings, familial and medical history, lab tests, medication, etc) represented in the EHR4CR Common Data Elements and the EHR4CR information model.

The expressiveness of the EHR4CR semantic interoperability frameworks is contingent on the coverage of both the patient information models (EHR4CR information model and Common Data Elements) and the clinical terminologies (integrated in the EHR4CR terminology) being used.

As stated figure 7, the EHR4CR terminology is built from both i) standards reference information models, templates and common data elements and ii) real-world constraints from pilot sites identified by the inventories (task 4.1b). We are currently formalizing 99 eligibility criteria from 10 clinical trials and the corresponding data elements and codes will also populate the EHR4CR Meta Data Repository and EHR4CR terminology (EHR4CR terminology v2.0).

In parallel, like in the epSOS project, we are currently mining CDA templates to extract data elements and their corresponding codes (codes of data elements, codes of values in value sets) to populate the EHR4CR Meta Data Repository and EHR4CR terminology (EHR4CR terminology v3.0).

### 3.2.2 Query model

Our purpose is to construct an explicit model of an EHR4CR query based on an extensible object model. This can be represented as a UML class diagram or suitable computable form that can be reasoned over for several purposes. A query model is distinct from an information model in that the former represents the questions that may be asked and the latter the facts that may be used to answer them. The query model is based on the HL7 RIM (on which the Abstract Information Model is also based) such that any query can be mapped onto the facts contained therein and will thus ensure that queries are computable.

In order to construct the query model using a methodical approach to meet the needs of EHR4CR use cases, an ad-hoc language ECLECTIC (Eligibility Criteria Language for European Clinical Trial Investigation and Construction) has been developed and is transformed into OCL queries.

### 3.3. Conclusion of the EHR4CR semantic interoperability framework

The EHR4CR semantic interoperability framework will be designed through an iterative process. The current components of the semantic interoperability framework described in the deliverable consists in the first version of the EHR4CR information model and the EHR4CR terminology that are currently evaluated in the context of the Protocol Feasibility use case of the EHR4CR project. The
expressiveness of the semantic interoperability framework is contingent on the coverage of the patient information models and of the clinical terminologies and being used. The results of the evaluation will be used to refine the EHR4CR information model both in its structure and scope. We are currently focusing on building the EHR4CR Common Data Elements that shall be stored in the EHR4CR Meta Data Repository.

The modeling efforts conducted as part of the WP4 shall be aligned with similar international initiatives. We have already identified for further cycles the need of a better integration of the modeling efforts done as part of the CDISC BRIDG and SHARE initiatives and the CEN 251 archetypes initiative. Especially, the EHR4CR Meta Data Repository shall be aligned to other MDRs (caDSR, openMDR, eMERGE repository, etc). These efforts will be conducted in collaboration with CDISC partner, as a contribution to the CDISC SHARE project. The semantic overlap of EHR4CR information model, metadata repository and terminology still need to be clearly defined in order to avoid replication of semantics and potential inconsistency.
4 References


6. Breil B, Semjonow A, Dugas M. HIS-based electronic documentation can significantly reduce the time from biopsy to final report for prostate tumours and supports quality management as well as clinical research. BMC Med Inform Decis Mak. 2009;9:5.


22. Niès J, Steichen O, Jaulent M-C. Archetypes as interface between patient data and a decision support system. AMIA Annual Symposium Proceedings. 2007;1060.