MIXHS

Clinical Data Integration Model

Core Interoperability Ontology for Research Using Primary Care Data

J.-F. Ethier[1]; V. Curcin[2]; A. Barton[1]; M. M. McGilchrist[3]; H. Bastiaens[4]; A. Andreasson[5]; J. Rossiter[6]; L. Zhao[6]; T. N. Arvanitis[6]; A. Taweel[7]; B. C. Delaney[8]; A. Burgun[1]

[1]INSERM UMR 1138 team 22 Centre de Recherche des Cordeliers, Faculté de médecine, Université Paris Descartes – Sorbonne Paris Cité, Paris, France;
[2]Department of Primary Care and Public Health, Imperial College London, London, United Kingdom;
[3]Public Health Sciences, University of Dundee, Dundee, United Kingdom;
[4]Department of Primary and Interdisciplinary Care, University of Antwerp, Antwerp, Belgium;
[5]Centre for Family Medicine, Karolinska Institute, Stockholm, Sweden and Stress Research Institute, Stockholm University, Stockholm, Sweden;
[6]Institute of Digital Healthcare, WMG, University of Warwick, Coventry, United Kingdom;
[7]Department of Informatics, King's College London, London, United Kingdom;
[8]NIHR Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London, London, United Kingdom

Keywords
Translational medical research, interoperability, phenotyping, ontology, primary care

Summary

Introduction: This article is part of the Focus Theme of *Methods of Information in Medicine* on "Managing Interoperability and Complexity in Health Systems".

Background: Primary care data is the single richest source of routine health care data. However its use, both in research and clinical work, often requires data from multiple clinical sites, clinical trials databases and registries. Data integration and interoperability are therefore of utmost importance.

Objectives: TRANSFoRm's general approach relies on a unified interoperability framework, described in a previous paper. We developed a core ontology for an interoperability framework based on data mediation. This article presents how such an ontology, the Clinical Data Integration Model (CDIM), can be designed to support, in conjunction with appropriate terminologies, biomedical data federation within TRANSFoRm, an EU FP7 project that aims to develop the digital infrastructure for a learning healthcare system in European Primary Care.

Methods: TRANSFoRm utilizes a unified structural/terminological interoperability framework, based on the local-as-view mediation paradigm. Such an approach mandates the global information model to describe the domain of interest independently of the data sources to be explored. Following a requirement analysis process, no ontology focusing on primary care research was identified and, thus we designed a realist ontology based on Basic Formal Ontology to support our framework in collaboration with various terminologies used in primary care.

Results: The resulting ontology has 549 classes and 82 object properties and is used to support data integration for TRANSFoRm's use cases. Concepts identified by researchers were successfully expressed in queries using CDIM and pertinent terminologies. As an example, we illustrate how, in TRANSFoRm, the Query Formulation Workbench can capture eligibility criteria in a computable representation, which is based on CDIM.

Conclusion: A unified mediation approach to semantic interoperability provides a flexible and extensible framework for all types of interaction between health record systems and research systems. CDIM, as core ontology of such an approach, enables simplicity and consistency of design across the heterogeneous software landscape and can support the specific needs of EHR-driven phenotyping research using primary care data.

Correspondence to:
Jean-François Ethier
INSERM UMR_S 872 team 22
Information Sciences to support Personalized Medicine
Centre de Recherche des Cordeliers
Rue de l'Ecole de Médecine
75006 Paris
France
E-mail: ethierj@gmail.com

Methods Inf Med 2015; 54: 16–23
http://dx.doi.org/10.3414/ME13-02-0024
received: June 14, 2013
accepted: April 23, 2014
Epub ahead of print: June 18, 2014

1. Introduction

Primary care data is the single richest source of routinely collected health care data. However its use, both in research and clinical work, often requires data from multiple clinical sites with different health record systems and integration with clinical trial and other types of medical data [1]. Data interoperability is therefore of utmost importance, and is typically implemented using a set of models and mappings [2]. There have been attempts to create generic information models to serve as standards, including the OpenEHR reference model,

the HL7 Reference Information Model (RIM) and the Clinical Data Acquisition Standards Harmonization (CDASH) model [3–7]. An ongoing international collaboration between standards organizations and industry partners, the Clinical Information Modeling Initiative (CIMI), aims at bringing together a variety of approaches to clinical data modeling (HL7 templates, openEHR archetypes, etc.) as a series of underlying reference models [8]. Nevertheless, many existing data sources are not designed according to these initiatives [9].

TRANSFoRm is an EU FP7 project that aims to comprehensively support the integration of clinical and translational research data in the primary care domain as part of a learning healthcare system [10, 11]. Its vision is demonstrated through three use cases: a genotype-phenotype around type 2 diabetes, a randomized clinical trial of treatment for gastroesophageal reflux disease and a diagnostic decision support system. It relies on software tools, such as a Query Formulation Workbench, a Study Manager and a Decision Support Ontological Evidence Service. They all need to access heterogeneous data sources. Moreover, the last two require the possibility of returning collected data back to the electronic health record (EHR) system. To that goal, the Clinical Data Integration Model (CDIM) was designed as the integration cornerstone for the project to enable interoperability between different types of data sources and different countries.

The *mediation* approach employed by CDIM allows structurally heterogeneous local sources to be used in distributed infrastructures [12]. A central information model is related to each local model via mappings. Queries are first expressed according to the central model and then "translated" by the system for each local source. Each source therefore retains its structure and control over its data. BIRN, caBIG and Advancing Clinico-Genomic Trials piloted this approach in the biomedical domain [13–15]. CDIM is the first mediation approach for primary care research.

Other approaches have been explored. One strategy relies on creation and maintenance of a *data warehouse,* to which data from each local data source is transferred. If the local source does not share the structure of the data warehouse, an Extract-Transform-Load (ETL) process is used to transfer and transform the data into the target structure. The i2b2 initiative is an example of such an approach [16]. A uniform and unique structure can then be used for queries. When local sources share a similar structure, *data federation* can be used, whereby instead of transferring data, queries are executed locally at source and the results aggregated. The ePCRN project explored this approach for primary care research, by ensuring the structure of all its sources conforms to the American Society for Testing and Materials Continuity of Care Record (CCR) information model [17, 18]. The Shared Health Research Information Network (SHRINE) uses a similar approach to federate i2b2 sources [19]. However, since TRANSFoRm has no control over the data sources' structure and since sources will not allow TRANSFoRm to use ETL, these approaches could not meet our requirements.

2. Objectives

TRANSFoRm's general approach relies on a unified interoperability framework, described in a previous paper [20]. We developed a core ontology for an interoperability framework based on data mediation. This article presents how such an ontology, the Clinical Data Integration Model, can be designed to support, in conjunction with appropriate terminologies, biomedical data federation within TRANSFoRm, an EU FP7 project that aims to develop the digital infrastructure for a learning healthcare system in European Primary Care.

3. Methods

The Clinical Data Integration Model (CDIM) was designed to represent clinical elements relevant to primary care and serve as a basis for data integration in the TRANSFoRm project. Data integration often relies on a combination of two types of models: information models (also called structural models) and terminological models (also referred to as semantic models). These two types of models, structural and terminological, are not independent as there are mutual constraints between the information models and coding systems [21] requiring these two models to be bound in order to fully assert their content [22].

For example, a field in a database might be named *dx* and contain the value *T90*. By binding the information model, where *dx* represents a patient diagnosis, with the terminological model used, the International Classification of Primary Care 2 (ICPC-2), we can assert that this represents a diagnosis of non-insulin dependent diabetes.[23,24] The equivalent representation using CDIM is achieved by binding the class *diagnosis (OGMS_0000073)*[a] with the term T90 from ICPC-2.

TRANSFoRm utilizes a unified structural/terminological interoperability framework, based on the local-as-view paradigm bringing together information models, terminologies and binding information, as shown in ▶Figure 1 [20]. The generic data queries expressed with CDIM are mapped to the local Data Source Model (DSM), so that they can be executed. Such an approach mandates the global information model to describe the domain of interest, independently of the data sources to be explored. ▶Figure 2 illustrates how the different models can interact together, through an example using the General Practice Research Database (GPRD) and the NIVEL Primary Care Database (NPCD) as data sources [25, 26].

Both types of models are required (information and terminological) since they each carry unique types of information. Terminologies express generic concepts of disease or state without implying the clinical context in which the data is created or used [27, 28]. The same concept can be used to represent a possible diagnosis, a confirmed diagnosis, or a comorbidity. It

[a] Throughout the text, ontology classes and properties will be italicized with RDF identifiers presented in parentheses. Here, the class diagnosis bears the rdf:id OGMS_0000073 since the class is imported in CDIM from the Ontology of General Medical Science.

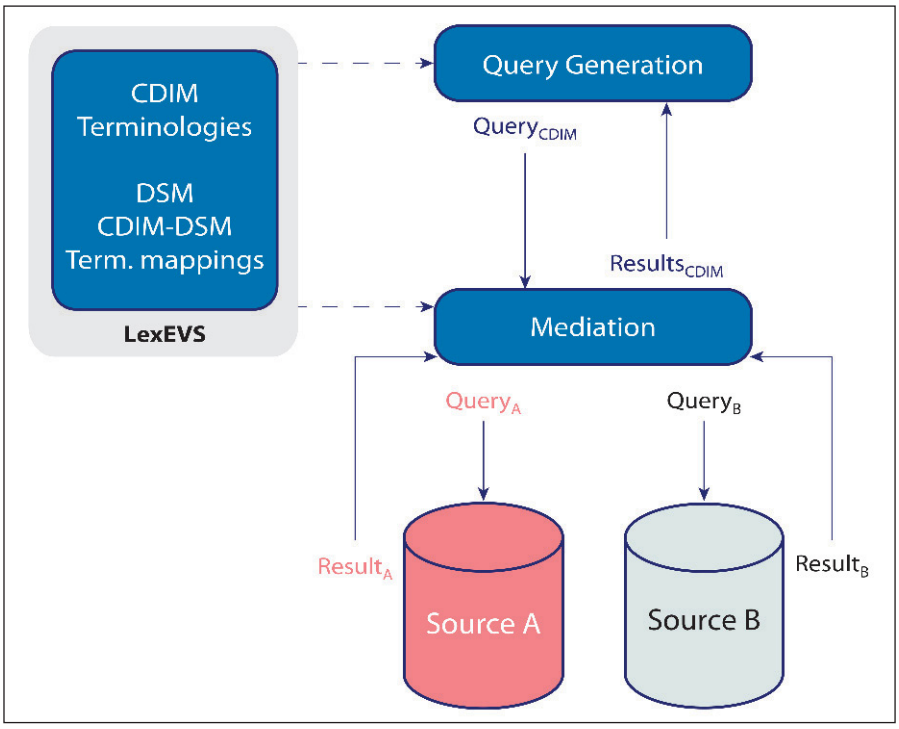

Figure 1 Interoperability framework based on CDIM in context of the Query Formulation Workbench. CDIM and terminologies are bound together to express queries independently of specific sources. Data source models (DSM), CDIM to DSM mappings and terminology mappings are used to translate the query during the mediation stage in order to execute it on local sources and provide unified results.

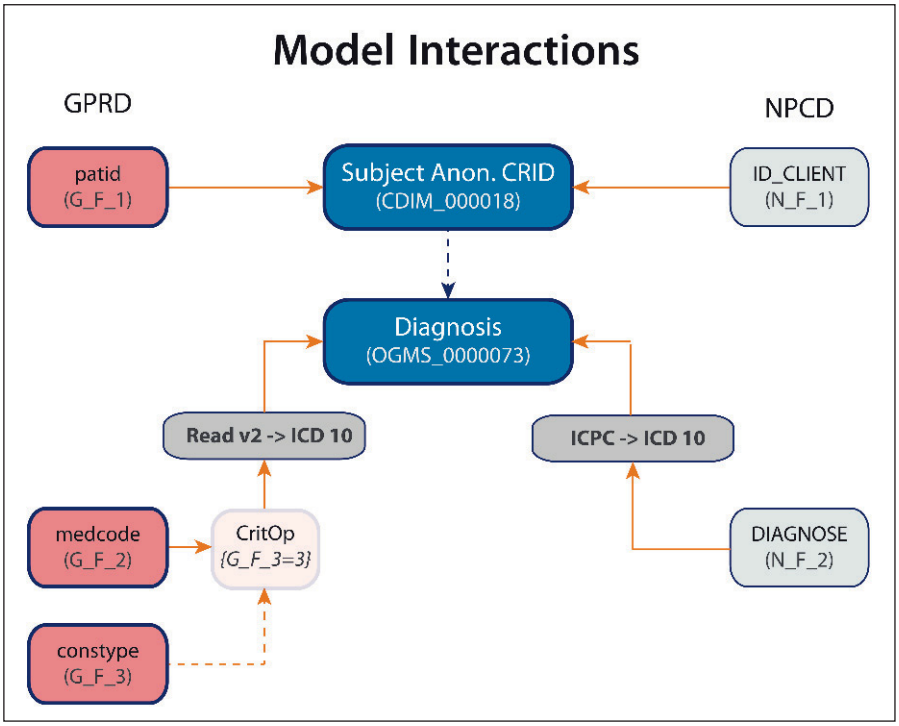

Figure 2 Model interactions in the task of retrieving a list of patient identifiers and diagnoses. In the GPRD database, the "medcode" field contains a diagnosis only if the field "constype" is equal to 3 for the same record. (DSM: GPRD and NPCD; blue boxes – CDIM classes, grey boxes – terminological mappings).

may also be used in the history section of a patient record to represent a problem that occurred years before or even in the patient's family. Moreover, a terminology like the International Classification of Disease 10 (ICD 10) is meant to be used by various systems (e.g., public health surveillance, electronic health records, billing systems) [29, 30].

On the other hand, information models usually focus on high level concepts (e.g., diagnosis) and omit particular representations of data (e.g., adenocarcinoma of the prostate), in order to be flexible and support binding with multiple terminologies, which might vary in depth and coverage. Furthermore, they provide the structure that is used to organize patient data in health records and databases (structural models).

Nevertheless, there is a grey zone where certain concepts might be found both in information models and terminologies. For example, should an information model contain concepts like *Type 1 diabetes mellitus* and *Type 2 diabetes mellitus*? Or, should it only contain the concept diagnosis, and rely on terminologies to support the relationships between these two diabetes concepts, as they can also be found in ICD-10-CM for example (codes *E10* and *E11*)? This underlines the importance of recognizing that information models and terminologies are not discrete entities, but rather a continuum along which the appropriate abstractions are constructed.

When developing CDIM, if some information was to be found in a recognized terminology (e.g., diabetes concepts are present in the ICPC-2 and ICD-10-CM), then only the "parent" concept was included in CDIM (e.g., *Disease*). However, exceptions were occasionally made for efficiency purposes, when a concept would frequently appear in queries. Taking blood pressure as an example, a systolic blood pressure measurement of 100 mmHg could be expressed with two triplets, linked together:

- *physical examination* = systolic blood pressure measurement
- *measurement datum* = 100 mmHg.

Yet, if included in CDIM, its expression only requires the assignments:

- *systolic blood pressure measurement datum* = 100 mmHg.

Given the extensive use of such measurements, including it in CDIM simplifies query construction.

3.1 Content Development

The specific requirements for primary care data were first gathered through discussions with experts in the field, as well as, through a sampling of various research criteria in order to get a broad view of the domain [31]. The continuum of primary care aims at following the patients from birth to death, including disease treatment and preventive care. As opposed to specialist care, primary care data tend to include longer follow-up time and a broader view of the patient, but with less detailed information. The primary care patient population reflects all degrees of disease severity and co-morbidity compared to disease specific records where sub-populations are followed. The particular nature of primary care data makes it especially well-suited to support "real-world" evaluations or to study care trajectories [32].

Although many clinical concepts such as diagnosis, medication or demographics are not unique to primary care, two are specifically important in primary care: *reason for health care encounter* and *health care episode*. The former captures the fact that patients often seek medical attention because of a sign or symptom that may or may not eventually lead to an established diagnosis. Within CDIM, *Reason for health care encounter* is represented as a role, in order to enable both symptoms and diseases to be qualified as the main reason for the visit [33]. For example *abdominal pain* would have the role *reason for health care encounter role* during the initial visit, and *Crohn's disease* or *pancreatitis* could hold this role in subsequent encounters.

The *health care episode* (often referred to as "episode of care") is introduced to take into account the fact that patients will often see their primary care physician for longitudinal follow-up. As a result, although multiple encounters might be coded with a diagnosis of *major depression*, they might all be related to the same *major depression*. Furthermore, the diagnostic problem may evolve during an episode of care as new information is gathered. Recognizing this is crucial to proper assessment of incidence and related measures [34]. CDIM captures this semantic using the class *health care episode,* with the axiom "*health care episode has_part* some *health care encounter*". A single encounter can then also be part of multiple *health care episodes* as multiple problems can be addressed during one visit.

In order to address the integrative requirements of primary care data, CDIM also contains organizational concepts, such as physical practices. In TRANSFoRm, this allows CDIM queries to refer to a specific set of practices as selected by the researcher. Supporting organizational units in CDIM also allows a more finely grained control over data access security, as policies can be applied distinctly to different subsets of data.

Genetic technology is rapidly evolving and the availability of genetic data is increasing, introducing new research questions and paradigms. Masys et al consider requirements for levels of integration of genomic data into Electronic Health Records [35]. Following this approach, CDIM supports interpretive codes which can be readily used for automated processes for single nucleotide polymorphisms (SNP), but not the full sequence information.

3.2 Ontology

CDIM supports the unification of structural, terminological and binding information. Traditionally, these models have been dealt with separately but they are interdependent and share requirements [21]. In order to address this interdependence and facilitate the framework's design and deployment, a decision was made to bring them together within one structure, and to rely on Mayo Clinic's LexEVS open-source terminology server as the storage solution, given its versatility and ability to handle multiple custom models, including ontologies [36].

As a mediation schema, CDIM needs to support data integration from multiple types of data sources. Current data sources used in TRANSFoRm include relational and XML databases, both standards based, such as HL7 CDA and non-standard ones, so the current interoperability framework is designed to support this [37].

Two general approaches exist in terms of formal ontologies: the realist and the cognitivist approaches. A cognitivist ontology aims at formalizing the concepts we use to categorize the world, as revealed by our common sense and our language: such an ontology has a cognitive and linguistic bias. For example, the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) categories are thought of as cognitive artifacts, ultimately depending on human perception, cultural imprints and social conventions [38, 39]. On the opposite side of the spectrum, a realist ontology aims at formalizing the real entities of the world, which we know through our best scientific theories [40]. In the biomedical domain, the OBO Foundry collection of interoperable ontologies is built upon the realist upper ontology BFO[41, 42]. The medical domain is seemingly a better fit for a cognitivist ontology, since it includes informational objects and mental constructs, such as diagnoses. However, these can also be efficiently formalized with a realist approach, as illustrated by the Ontology for General Medical Science (OGMS) [43], which formalizes a diagnosis as an informational content entity about the health status of a patient.

CDIM was designed as a realist ontology and uses Basic Formal Ontology (BFO) 1.1 as the foundational ontology [44], based on BFO's central role in the OBO Foundry. Several OBO Foundry ontologies, including OGMS, the Vital Sign Ontology (VSO) and the Information Artifact Ontology (IAO) were directly imported into CDIM [45, 46]. CDIM also integrates classes from other ontologies such as the Ontology for Biomedical Investigations (OBI) and the Gene Ontology (GO) [47, 48].

4. Results

CDIM introduced over 100 new classes and several additional properties and axioms, which in combination with imported ontologies resulted in the total of

549 classes and 82 properties. As CDIM is stored inside a LexEVS instance, all imports are merged into a single .owl file, created directly in Protégé through the Refactor/Merge ontologies tool, enabling easier load processing in the framework.

Temporal aspects are rarely, if at all, covered in the existing ontologies that we imported. As these play a crucial role in defining clinical eligibility criteria, we created 25 new classes to express these concepts. Whenever possible, we relied on equivalent classes instead of using anonymous classes, in order to support operations not based on Semantic Web reasoning techniques. Equivalent classes provide URIs that are then used as mapping targets for the CDIM-DSM mapping models. Internal validity and consistency was checked using the semantic reasoner HermiT 1.3.8 [49].

CDIM design also required addition of some axioms to imported classes. For example, a diagnostic process can take a long time before completion and production of a diagnosis. It is therefore important to identify the end of the process, in order to correctly attach temporal information to the resulting diagnosis. This temporal aspect is currently lacking in OGMS, therefore we added the following classes and axioms in CDIM:

- The equivalent class *diagnostic process conclusion instant* defined as
  - "*temporal_instant* and (*has temporal occupant* some *diagnostic process conclusion*)"
- The class *diagnostic process conclusion* was created and defined as:
  - a subclass of the BFO *process boundary* class.
  - it also bears the axiom "*occupies temporal region* some *diagnostic process conclusion instant*", linking it to the temporal information.
- The *diagnostic process* class was enriched by adding the axiom:
  - "*ends_with* some *diagnostic process conclusion*" to define its final subprocess as *diagnostic process conclusion.*

CDIM was evaluated in terms of its capacity to support query definitions required by the three use cases in TRANSFoRm. The first one is an epidemiological study on genetic risk markers for response to treatment in diabetes mellitus type 2. The main question is "Are well selected single nucleotide polymorphisms (SNPs) in type 2 diabetic patients associated with variations in drug response to oral antidiabetics (Sulfonylureas)?"[50]. The second use case is a randomized controlled trial investigating on-demand vs. continuous use of proton pump inhibitors (PPIs) in treating gastroesophageal reflux disease (GORD) and its impact on symptom relief and quality of life in patients [51]. Finally, the third use case consists of evaluating approaches to provide diagnostic decision support, based on existing EHR data, reason for encounter and captured clinical clues.

One of the major requirements for the first two use cases in TRANSFoRm is the ability to identify eligible patients in EHRs and other primary care data sources. Previous research found that two thirds of all information needed to assess the eligibility of a patient for a trial are related to disease history, namely disease, symptoms, signs and diagnostic or lab tests, and treatment history [31], which also applies to the TRANSFoRm use cases. One of the crucial aspects is to minimize misclassification, while identifying eligible patients. As also found by the eMerge project, it is important to not solely rely on diagnostic codes to identify diagnoses, but to also use other patient characteristics like laboratory tests or medication to verify the diagnosis [52–54].

The data elements needed for these studies were described in detail by the project's clinical researchers to ensure concepts coverage in CDIM. The main clinical concepts were: diagnoses (recent and medical history), laboratory tests, technical investigations (upper endoscopy), medications, symptoms and signs (difficulties swallowing, signs of gastrointestinal bleeding, unintentional weight loss), physical examination data (blood pressure, weight, height). The genetic concepts needed for the diabetes use case could be limited to SNPs. The following information also needed to be provided: moment of diagnosis; dates, values and units for all measurements; dates, number and dose for medication.

For example, a *formulated pharmaceutical* can be characterized through several data item entities, including *active ingredient data item*, *dose form data item* and *strength data item*. Such formalization can be made compatible with pre-existing norms – for example, RxNorm's category *semantic clinical drug form* could be formalized as the association of CDIM classes *active ingredient data item* and *dose form data item* [55]. Additionally, the instruction given by a prescription can be formalized as a subclass of OBI's *directive information entity*, composed of several *directive information entity parts*. For example, the prescription "take Metformin 500 mg 3 times a day during two weeks" is composed of "3 times a day" (which is an instance of *administration frequency item*) and "during two weeks" (which is an instance of *duration of treatment period item*).

Some items from the use cases have not been included in CDIM. These were the ones mostly focusing on habits (e.g., level of physical activity/sedentarism, dietary habits) or behavioral interventions (e.g., status of self-management education, or performance of self-measurement of blood glucose). Although very important concepts, they were deemed too specific to a research area or very rarely encountered in current data sources. CDIM usage will be regularly reviewed to inform future classes additions and deletions.

All concepts identified, as required by researchers, were successfully expressed in queries using CDIM and terminologies. We shall now present an example of how triplets using CDIM classes, operators and terminologies (or values) can be created and used in TRANSFoRm tools.

4.1 Application to TRANSFoRm

The TRANSFoRm Query Formulation Workbench provides a user interface for clinical researchers to create clinical studies, design eligibility criteria, initiate distributed queries, monitor query progress, and report query results. It captures eligibility criteria in a computable representation, which is based on CDIM ontology so the criteria can be translated into executable query statements on the data source side using CDIM to data source model mappings. They are then grouped to form application friendly reusable units.

Figure 3 Workbench criteria editor uses CDIM classes to create queries which can be applied to multiple primary care databases used by the TRANSFoRm project.

Let us consider an example inclusion criterion for patients who had an HbA1c test result of ≥ 6.5% on or before the 16/04/2013 date. The Laboratory Measurement group aggregates relevant concepts closely related to the laboratory test class extracted from CDIM, such as test type, date of test and test value. It is one of seven categories (like demographics, medications, etc.) currently used within the Workbench. The structure allows new categories to be easily added as per user requirements.

The example criterion is specified by a user of the Query Formulation Workbench, as shown in ▶Figure 3. The Laboratory Test artifact is presented to the user in the form of a template for entering values for operators and values. Resulting triplets would be:

- *laboratory_Test_Type_ID* = [LOINC; 4548 – 4] [b]
- *laboratory_measurement_datum* ≥ 6.5
- *laboratory_measurement_unit_label* = [UO; 0000187] [c]
- *lab_result_confirmation_instant* ≤ 2013/04/16

A query expressed in this way is passed to the data source, where a translation component uses CDIM (and its mappings to the local source model) to convert the query into a representation understandable by the local data source and extract results to send back to the researcher.

5. Discussion

TRANSFoRm is one of several complementary initiatives that develop services and tools to foster more efficient research using EHR data. Furthermore, only in aligning primary and secondary/tertiary care data can a full picture of patient's clinical evolution be constructed. Therefore, facilitating interoperability between TRANSFoRm and other initiatives is essential. Using an ontology, as the core model, allows for formal logic to be used to define classes and their relationships, promoting a shared, well defined view of a domain. It is possible to reason about data elements present over multiple sources, and define new relationships.

Specific classes, such as *reason for health care encounter* or *health care episode*, were designed in such a way as to avoid inconsistencies with other common classes. For example, a *reason for health care encounter* was formalized as an entity bearing a special role that we called the *reason for health care encounter role*. Thus, it was not necessary to modify the class *diagnosis*, *symptom* and *sign* in our ontology so that they could be a *reason for health care encounter*, as all these entities can bear the *reason for health care encounter role*. Therefore, the CDIM approach can reuse both existing terminologies (e.g., ICPC-2) and increasingly popular semantic web resources such as SPARQL repositories [58].

The reusability of CDIM is thereby enhanced since a large part of the specific requirements can be handled by the binding of terminologies providing sufficient precision and coverage for the desired context [59]. This facilitates ontology alignment and interoperability with other projects using ontologies, such as epSOS, that aims to develop a cross-border electronic health information transfer and also relies on BFO [60].

An additional benefit is that the necessary references to ontologies and terminologies can be created and embedded within existing standards, such as the CDISC Operational Data Model, which do not necessarily support ontologies, least of all the native creation of complex data elements constrained against both a clinical and research ontology [61]. Systems that support standards can thus be rapidly extended to support CDIM, without abandoning the existing standard.

The CDIM ontology is by definition extensible, but the question arises as to what extent CDIM should be extended as new concepts are required, or leave this to the terminology. It is to be expected that not every single point, possibly evaluated in a research project, will make its way in CDIM. Some niche concepts might never be included, in order to keep the ontology manageable and relevant to most users.

Nevertheless, extensively relying on terminologies does imply that the project has much less control on content and definition of concepts. For example, the ICPC-2 classifies diabetes as insulin dependent (*T89*) and non-insulin dependent (*T90*) diabetes. This has been revised and current

[b] Logical Observation Identifiers Names and Codes (LOINC) is a universal code system for identifying laboratory and clinical observations and the HbA1c test is represented by the code 4548–4 [56]

[c] Units are represented as Ontology of Units of Measurements (UO). The unit for HbA1c is % (ratio), with UO code value 0000187 [57]

approaches use mainly type-1 and type-2. Equivalences between these terms are not perfect as some type 2 "depend" on insulin for their treatment. However, this reflects the state of limitations for existing data. When an equivalence does exist between concepts, terminological heterogeneity can be mended by using inter-terminology mappings like those offered by the UMLS [62].

Of note, the local-as-view mediation approach mandates that the decision to include a concept or not must be based on relevance to the users and not to its availability (or not) in data sources: a concept useful for many queries will be included even if no current data source contains it. In this context, a high number of queries using a concept but returning no data is highly informative. As incentives are put in place to foster the use of EHRs, such information might help focus such incentives in terms of research priorities.

TRANSFoRm uses data and process provenance, as a means to achieve traceability and auditability in its digital infrastructure. The novelty of the TRANSFoRm provenance framework is that it links the provenance model, represented with the Open Provenance Model standard, to the medical domain models, by means of bridging ontologies, thereby enabling verification with respect to established concepts [63]. CDIM is a key element in this approach, since it allows a uniform conceptualization of annotations in provenance traces that are produced by multiple tools and across national boundaries. This has direct impact on the ability of the system to be audited in a consistent manner regardless of its geographical location, e.g., a clinical trial design conducted in Germany, or a record of data extraction for an epidemiological study in France.

Genetic (and eventually proteomic and metabolic) primary observations will be more easily leveraged with time and as their availability increases. At some point, sequence structural variations and mutations, as well as, gene expression data will be relevant to the researcher and such concepts will also need to be included in CDIM. Nevertheless it is unclear, given the high heterogeneity inherent to the field of translational research, and the increasing use of genetic information in personalized medicine to which level of precision the models will need to abide by.

6. Conclusion

A unified mediation approach to semantic interoperability provides an extensible framework for interactions between health record systems and research systems. CDIM, as a core ontology of such an approach, enables simplicity and consistency of design across the heterogeneous software landscape and can support the specific needs of EHR-driven phenotyping, using primary care data. This was demonstrated in TRANSFoRm, where the software tools such as the Query Workbench are agnostic of the structural and terminological details of the data sources they interact with.

CDIM is flexible and modular by design as it can be bound to multiple terminologies, enabling new ways to approach data as the requirements of translational medicine evolve and new domains like epigenetic become part of patient care.

Acknowledgments

We would like to thank our colleagues from the TRANSFoRm project for their support and insightful discussions regarding our endeavor. This work was partially supported by the European Commission Framework 7 Programme – DG INFSO (FP7 247787). This research is also partially supported by the National Institute for Health Research (NIHR) Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London, NHS England for the Institute of Digital Healthcare at WMG, University of Warwick and the Institut National de la Santé et de la Recherche Médicale (INSERM).

References

1. Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. BMJ 2003; 326: 1070.
2. Sujansky W. Heterogeneous Database Integration in Biomedicine. J Biomed Inform 2001;34:285–98.
3. Beale T, Heard S, Kalra D, et al. The openEHR Reference Model – EHR Information Model – Release 1.0.2 [Internet]. 2008 [cited 2012 Jun 29].Available from: http://www.openehr.org/releases/1.0.2
4. Murphy SN, Mendis M, Hackett K, et al. Architecture of the Open-source Clinical Research Chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc 2007. pp 548–552.
5. Schadow G, Mead CN, Walker DM. The HL7 reference information model under scrutiny. Stud Health Technol Inform 2006; 124: 151–156.
6. CDASH – Basic Recommended Data Collection Fields for Medical Research [Internet] [cited 2012 Dec 8]. Available from: http://www.cdisc.org/cdash
7. López DM, Blobel B. Architectural approaches for HL7-based health information systems implementation. Methods Inf Med 2010; 49: 196 –204.
8. Clinical Information Modelling Initiative... [Internet]. [cited 2012 Dec 8]. Available from: http://www.openehr.org/326-OE.html?branch=1&language=1
9. Ohmann C, Kuchinke W. Future developments of medical informatics from the viewpoint of networked clinical research. Interoperability and integration. Methods Inf Med 2009; 48: 45 –54.
10. Delaney B. TRANSFoRm: Translational Medicine and Patient Safety in Europe. In: Grossman C, Powers B, McGinnis JM, editors. Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary. Washington, DC: National Academies Press; 2011. pp 198–202.
11. TRANSFoRm Project [Internet]. [cited 2012 Apr 11];Available from: http://www.transformproject.eu
12. Wiederhold G. Mediators in the architecture of future information systems. Comput J 1992; 25: 38 – 49.
13. Gupta A, Ludascher B, Martone ME. BIRN-M: a semantic mediator for solving real-world neuroscience problems. In: Halevy AY, Ives ZG, Doan A, editors. Proc ACM SIGMOD Int Conf Manag Data. New York, NY: ACM Press; 2003. pp 678– 678.
14. Stanford J, Mikula R. A model for online collaborative cancer research: report of the NCI caBIG project. Int J Healthc Technol Manag 2008; 9: 231–246.
15. Martin L, Anguita A, Graf N, et al. ACGT: advancing clinico-genomic trials on cancer – four years of experience. Stud Health Technol Inform 2011; 169: 734 –738.
16. Murphy SN, Weber G, Mendis M, et al. Serving the Enterprise and Beyond with Informatics for Integrating Biology and the Bedside (i2b2). J Am Med Inform Assoc 2010; 17: 124 –130.
17. Delaney BC, Peterson KA, Speedie S, et al. Envisioning a Learning Health Care System: The Electronic Primary Care Research Network, A Case Study. Ann Fam Med 2012; 10: 54 –59.
18. Peterson KA, Fontaine P, Speedie S. The Electronic Primary Care Research Network (ePCRN): A New Era in Practice-based Research. J Am Board Fam Med 2006; 19: 93–97.
19. Weber GM, Murphy SN, McMurry AJ, et al. The Shared Health Research Information Network

(SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories. J Am Med Inform Assoc 2009; 16: 624–630.
20. Ethier J-F, Dameron O, Curcin V, et al. A unified structural/terminological interoperability framework based on LexEVS: application to TRANSFoRm. J Am Med Inform Assoc 2013; Published Online First: April 9, 2013.
21. Qamar R, Kola JS, Rector AL. Unambiguous data modeling to ensure higher accuracy term binding to clinical terminologies. AMIA Annu Symp Proc 2007. pp 608–613.
22. Rector AL. Clinical terminology: why is it so hard? Methods Inf Med 1999; 38: 239–252.
23. WHO|International Classification of Primary Care, Second edition (ICPC-2) [Internet]. WHO. [cited 2013 Jun 13]. Available from: http://www.who.int/classifications/icd/adaptations/icpc2/en/
24. Rector AL. Thesauri and formal classifications: terminologies for people and machines. Methods Inf Med 1998; 37: 501–509.
25. Clinical Practice Research Datalink – CPRD [Internet]. [cited 2012 Jul 28]. Available from: http://www.cprd.com/intro.asp
26. NIVEL|LINH [Internet]. [cited 2012 Jul 28]. Available from: http://www.nivel.nl/en/netherlands-information-network-general-practice-linh
27. Chute CG, Elkin PL, Sherertz DD, et al. Desiderata for a clinical terminology server. Proc AMIA Symp 1999. pp 42–46.
28. Cimino JJ. Terminology tools: state of the art and practical lessons. Methods Inf Med 2001; 40: 298–306.
29. WHO|International Classification of Diseases (ICD) [Internet]. WHO. [cited 2013 Jun 13]. Available from: http://www.who.int/classifications/icd/en/
30. Prins H, Hasman A. Appropriateness of ICD-coded diagnostic inpatient hospital discharge data for medical practice assessment. A systematic review. Methods Inf Med 2013; 52: 3–17.
31. Köpcke F, Trinczek B, Majeed RW, et al. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC Med Inform Decis Mak 2013; 13: 37.
32. De Lusignan S, Pearce C, Shaw NT, et al. What are the barriers to conducting international research using routinely collected primary care data? Stud Health Technol Inform 2011; 165: 135–140.
33. Arp R, Smith B. Function, role, and disposition in basic formal ontology. Nat Preceedings 2008; 1–4.
34. Soler JK, Okkes I, Oskam S, et al. Revisiting the concept of "chronic disease" from the perspective of the episode of care model. Does the ratio of incidence to prevalence rate help us to define a problem as chronic? Inform Prim Care 2012; 20: 13–23.
35. Masys DR, Jarvik GP, Abernethy NF, et al. Technical desiderata for the integration of genomic data into Electronic Health Records. J Biomed Inform 2012; 45: 419–422.
36. LexEVS 6.0 Architecture [Internet]. [cited 2013 May 30]. Available from: https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexEVS_6.0_Architecture
37. Health Level Seven International – Homepage [Internet]. [cited 2013 Jun 13]. Available from: http://www.hl7.org/
38. Grenon pierre. Bfo in a nutshell: A bi-categorical axiomatization of bfo and comparison with dolce [Internet]. University of Leipzig; 2003 [cited 2013 Jun 13]. Available from: www.ifomis.org/Research/IFOMISReports/IFOMIS Report 06_2003.pdf
39. Gangemi A, Guarino N, Masolo C, et al. Sweetening Ontologies with DOLCE [Internet]. In: Gómez-Pérez A, Benjamins VR, editors. Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web. Springer Berlin Heidelberg; 2002 [cited 2013 Dec 15]. pp 166–181. Available from: http://link.springer.com/chapter/10.1007/3-540-45810-7_18
40. Grenon P, Smith B. SNAP and SPAN: Towards Dynamic Spatial Ontology. Spat Cogn Comput 2004; 4: 69–104.
41. Smith B, Ashburner M, Rosse C, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007; 25: 1251–1255.
42. Smith B, Brochhausen M. Putting biomedical ontologies to work. Methods Inf Med 2010; 49: 135–140.
43. Scheuermann RH, Ceusters W, Smith B. Toward an Ontological Treatment of Disease and Diagnosis. AMIA Summit Transl Bioinforma 2009. pp 116–120.
44. Grenon P, Smith B, Goldberg L. Biodynamic ontology: applying BFO in the biomedical domain. Stud Health Technol Inform 2004; 102: 20–38.
45. Goldfain A, Smith B, Arabandi S, et al. Vital Sign Ontology. In: Proceedings of the Workshop on Bio-Ontologies. Vienna: 2011. pp 71–74.
46. The Information Artifact Ontology (IAO) is an ontology of information entities based on the BFO [Internet]. [cited 2012 Dec 9]. Available from: http://code.google.com/p/information-artifact-ontology/
47. Brinkman RR, Courtot M, Derom D, et al. Modeling biomedical experimental processes with OBI. J Biomed Semant 2010; 1: S7.
48. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25: 25–29.
49. Shearer R, Motik B, Horrocks I. HermiT: A highly-efficient OWL reasoner. In: Proceedings of the 5th International Workshop on OWL: Experiences and Directions (OWLED 2008) 2008. pp 26–27.
50. Pearson ER, Donnelly LA, Kimber C, et al. Variation in TCF7L2 influences therapeutic response to sulfonylureas: a GoDARTs study. Diabetes 2007; 56: 2178–2182.
51. Leysen P, Bastiaens H, Van Royen P. TRANSFoRm: Development of Use Cases [Internet]. [cited 2013 Feb 28]. Available from: http://transformproject.eu/Deliverable_List_files/D1.1%20Detailed%20Use%20Cases_V2.1–2.pdf
52. De Lusignan S, Khunti K, Belsey J, et al. A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: a pilot and validation study of routinely collected data. Diabet Med J Br Diabet Assoc 2010; 27: 203–209.
53. Manuel DG, Rosella LC, Stukel TA. Importance of accurately identifying disease in studies using electronic health records. BMJ 2010; 341: c4226.
54. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 2013; 20: e147–e154.
55. Nelson SJ, Zeng K, Kilbourne J, et al. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc JAMIA 2011; 18: 441–448.
56. Logical Observation Identifiers Names and Codes (LOINC®) — LOINC [Internet]. [cited 2013 Jun 13]. Available from: http://loinc.org/
57. unit-ontology – Ontology of Units of Measurement [Internet]. [cited 2013 Jun 13]. Available from: http://code.google.com/p/unit-ontology/
58. García Godoy MJ, López-Camacho E, Navas-Delgado I, et al. Sharing and executing linked data queries in a collaborative environment. Bioinforma Oxf Engl 2013;
59. Cimino JJ. High-quality, standard, controlled healthcare terminologies come of age. Methods Inf Med 2011; 50: 101–104.
60. epSOS: About epSOS [Internet]. [cited 2012 Apr 11]. Available from: http://www.epsos.eu/home/about-epsos.html
61. Kuchinke W, Wiegelmann S, Verplancke P, et al. Extended Cooperation in Clinical Studies through Exchange of CDISC Metadota between Different Study Software Solutions. Methods Inf Med 2006; 45: 441.
62. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004; 32: D267–270.
63. Curcin V, Danger R, Kuchinke W, et al. Provenance Model for Randomized Controlled Trials. In: Liu Q, Bai Q, Giugni S, Williamson D, Taylor J, editors. Data Provenance and Data Management in eScience. Berlin Heidelberg: Springer; 2013. pp 3–33.