Integrating resources for translational research : a unified approach for learning health systems

Jean-Francois Ethier

To cite this version:

Jean-Francois Ethier. Integrating resources for translational research : a unified approach for learning health systems. Information Retrieval [cs.IR]. Université Pierre et Marie Curie - Paris VI, 2016. English. <NNT : 2016PA066078>. <tel-01375481>

HAL Id: tel-01375481
https://tel.archives-ouvertes.fr/tel-01375481

Submitted on 3 Nov 2016

Université Pierre et Marie Curie

ECOLE DOCTORALE PIERRE LOUIS DE SANTE PUBLIQUE A PARIS : EPIDEMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMEDICALE

INSERM 1138, équipe 22

Intégration de ressources en recherche translationnelle :

une approche unificatrice en support des systèmes de santé « apprenants »

Par Dr Jean-François Ethier

Thèse de doctorat

Spécialité : Informatique biomédicale

Dirigée par Dr Anita Burgun

Présentée et soutenue publiquement le 16 février 2016

Devant un jury composé de :

M. Alain VANASSE	Professeur	Rapporteur
M. Olivier BODENREIDER	Chercheur	Rapporteur
Mme Anita BURGUN	Professeur	Directrice
M. Brendan DELANEY	Professeur	Examinateur
M. Patrice DEGOULET	Professeur	Examinateur
M. Marc CUGGIA	Professeur	Examinateur
Mme. Brigitte SEROUSSI	Maitre de conférence	Examinateur

À Pascale,

mon amie, ma confidente, ma force.

Remerciements

Je tiens premièrement à remercier profondément Anita. Faire le saut du Québec vers Rennes pour explorer le domaine de l'informatique médicale était un saut dans le vide et j'ai eu cette chance immense de trouver sur ma route une collègue aussi sensationnelle. Avec force de patience et de confiance, elle a su me guider dans les aléas de l'apprentissage de la recherche et de la politique dans le domaine. J'ai toujours pu compter sur elle tout au long du chemin, dans les moments exaltants comme dans les périodes de remise en question. Je lui dois une grande partie du scientifique que je suis aujourd'hui et pour ceci, je lui en serai pour toujours redevable et reconnaissant.

Ce saut dans le vide, je l'ai fait en compagnie de la femme de ma vie, Pascale. Par une décision que certains jugeaient insensée, elle a accepté de m'accompagner en France pour mes études, laissant son travail d'enseignante au Québec de côté. Son abnégation, sa gentillesse, son amour et son aide pour les multiples révisions ont fait en sorte que j'ai pu poursuivre mon rêve. Le sacrifice que cela représente est effectivement énorme et j'espère pouvoir, dans l'avenir, lui remettre au centuple ce qu'elle m'a donné, car c'est certainement ce qu'elle mérite, et même plus. La seule façon de comprendre son choix est de le faire à l'aune de l'amour que nous partageons. Bien évidemment, je me dois de faire un gros câlin aussi à Anne-Charlotte, notre petite princesse de trois ans qui a accepté si souvent de voir partir papa en avion et d'être heureuse de me voir revenir par la suite. J'aimerais aussi mentionner ici ma mère, Danielle Côté, qui m'a initié dès le plus jeune âge à la rigueur et à l'importance du travail bien fait. Ces valeurs m'ont accompagné tout au long de mon cheminement académique et m'ont grandement aidé à former mon esprit critique.

Scientifiquement, j'ai débuté mon parcours d'informatique médicale à Rennes et j'ai eu le bonheur de travailler avec Olivier Dameron et Pascal Van Hille avec qui j'ai appris beaucoup. J'ai aussi eu la chance d'avoir une ange-gardienne, Delphine, qui a su gérer les complications administratives afin que je puisse me concentrer sur la science.

Ce travail s'inscrit en grande partie dans le projet européen TRANSFoRm. Anita m'a proposé d'y participer et ce fut une occasion en or de travailler avec un groupe dynamique, brillant et dévoué entièrement à l'avancement du domaine. Brendan Delaney et Vasa Curcin ont su apporter la vision, le leadership et le support nécessaire à la réalisation de ce projet ambitieux. Plus précisément, j'aimerais souligner la patience et l'approche pédagogique de Vasa dans le cadre de l'écriture des différents articles et communications scientifiques. Mon travail dans TRANSFoRm c'est effectué dans le groupe de travail de Mark McGilchrist, véritable figure de mentor. Il est devenu un précieux collègue avec qui développer mes idées et les raffiner. Ce fut un plaisir de travailler avec lui et ses bons conseils ont certainement bonifié mon approche. Une mention spéciale aussi pour Derek et Anna avec qui j'ai eu beaucoup de plaisir à travailler.

Une thèse s'inscrit dans un processus administratif complexe et l'aide que Florence Bordu et Lydie Martorana m'ont apportée fut des plus précieuses pour me permettre de franchir la ligne d'arrivée. De même, je veux souligner le soutien et la compréhension de mes collègues du service de médecine interne de l'Université de Sherbrooke. Je tiens, en dernier lieu, à remercier mes rapporteurs Olivier Bodenreider et Alain Vanasse ainsi que les membres du jury Brendan Delaney, Marc Cuggia, Patrice Degoulet et Brigitte Seroussi d'avoir accepté d'accorder du temps à la réalisation de mon projet malgré leurs horaires surchargés.

Au final, je me sens énormément fortuné d'avoir autant de gens dans ma vie qui m'aident, me supportent et me poussent vers l'avant. À tous, merci !

Résumé

Les systèmes de santé apprenants (SSA) présentent une approche complémentaire et émergente aux problèmes de la recherche translationnelle en couplant de près la provision des soins de santé, la recherche (prospective et rétrospective) ainsi que les activités de transfert des connaissances. Afin de permettre un flot d'informations cohérent et optimisé dans le système, ce dernier doit se doter d'une plateforme intégrée de partage de données. Le travail présenté ici vise à proposer une approche de partage de données unifiée pour les SSA.

Les trois grandes familles de mise à disposition des données (entrepôt de données, fédération de données et médiation de données) sont analysées en regard des exigences des SSA pour finalement retenir la médiation. La sémantique des informations cliniques disponibles dans les sources de données biomédicales est la résultante des connaissances des modèles structurels des sources (ex. les diagnostics des patients sont dans le champ X de la table Y), mais aussi des connaissances des modèles terminologiques utilisés pour coder l'information (ex. Classification Internationale des Maladies 10^{e} révision – CIM-10). La structure de la plateforme unifiée qui prend en compte cette interdépendance est décrite.

La plateforme a été implémentée et testée dans le cadre du projet TRANSFoRm, un projet européen financé par le « Seventh Framework Program for research, technological development and demonstration », qui vise à développer un SSA incluant les soins de première ligne. L'instanciation du modèle de médiation pour le projet TRANSFoRm, le Clinical Data Integration Model (CDIM), est présentée et analysée. Sont aussi présentés ici les résultats d'un des cas d'utilisation de TRANSFoRm en regard de la plateforme unifiée de données pour supporter la recherche prospective afin de donner un aperçu concret de l'impact de la plateforme sur le fonctionnement du SSA.

Au final, la plateforme unifiée de médiation proposée ici permet un niveau d'expressivité suffisant pour les besoins du SSA TRANSFoRm. Le système est flexible et modulaire et le modèle de médiation CDIM couvre les besoins exprimés pour le support des activités d'un SSA incluant les soins de première ligne.

Mots clé : recherche translationnelle, système de santé apprenant, interopérabilité, médiation de données, terminologie, ontologie, phénotypage, soins primaires, LexEVS

Abstract

Learning health systems (LHS) are gradually emerging and propose further solutions to translational research challenges by implementing close coupling of health care delivery, research (both retrospective and prospective) as well as knowledge transfer activities. To support coherent knowledge sharing across the system, it needs to rely on an integrated and efficient data integration platform. The framework and its theoretical foundations presented here aim at addressing this challenge.

Data integration approaches can be grouped according to three high level categories: data warehousing, data federation and data mediation. They are analysed in light of the requirements derived from LHS activities and data mediation emerges as the one most adapted for a LHS. The semantics of clinical data found in biomedical sources can only be fully and properly derived by taking into account, not only information from the structural models (e.g. patient's diagnoses can be found field X of table Y), but also terminological information (e.g. codes from the International Classification of Disease 10th revision – ICD 10) used to encode facts. The unified framework proposed here takes into account this reality.

The platform has been implemented and tested in the context of TRANSFoRm, a European project funded by Seventh Framework Program for research, technological development and demonstration. It aims at developing a LHS including clinical activities in primary care. The mediation model developed for the TRANSFoRm project, the Clinical Data Integration Model (CDIM), is presented and discussed. Results from one of the TRANSFoRm use cases are also presented. They illustrate how a unified data sharing platform can support and enhance prospective research activities in the context of a LHS.

In the end, the unified mediation framework presented here allows sufficient expressiveness for the TRANSFoRm LHS needs. It is flexible, modular and the CDIM mediation model supports the requirements of a primary care LHS.

Keywords: Translational research, learning health systems, interoperability, data mediation, terminology, ontology, phenotyping, primary care, LexEVS

Publications

Articles

Ethier J-F, Delaney BC, Curcin V, McGilchrist MM, Lim Choi Keung SN, Zhao L, Andreasson A, Brodka P, Radoslaw M, Mastellos N, Burgun A, Arvanitis TN. Implementation and validation of a standards-based approach to embedding clinical trial functionality in routine electronic health record system (submitted to JAMIA January 2016)

Delaney BC, Curcin V, Andreasson A, Arvanitis TN, Bastiaens H, Corrigan D, Ethier J-F, Kostopoulou O, Kuchinke W, McGilchrist M, van Royen P, Wagner P. Translational Medicine and Patient Safety in Europe: TRANSFoRm-Architecture for the Learning Health System in Europe. *Biomed Res Int*. 2015;2015:961526.

Ethier J-F, Curcin V, Barton A, McGilchrist MM, Bastiaens H, Andreasson A, Rossiter J, Zhao L, Arvanitis TN, Taweel A, Delaney BC, Burgun A. Clinical data integration model. Core interoperability ontology for research using primary care data. *Methods Inf Med.* 2015;54(1):16–23.

Ethier J-F, Dameron O, Curcin V, McGilchrist MM, Verheij RA, Arvanitis TN, Taweel A, Delaney BC, Burgun A. A unified structural/terminological interoperability framework based on LexEVS: application to TRANSFoRm. *J Am Med Inform Assoc*. 2013 Oct;20(5):986–94.

Communications

McGilchrist MM, Curcin V, Ethier J-F, Arvanitis TN, Delaney BC. "TRANSFoRm connectivity infrastructure to embed eCRFs into EHR systems", *In 2015 AMIA Clinical Research Informatics Summit*, San Francisco, CA: USA, March 2015

Lim Choi Keung SN, Zhao L, Ethier J-F, Curcin V, Burgun A, Delaney BC, Arvanitis TN. "Semantic Enrichment of CDISC Operational Data Model", Papers/Podium Presentations. *In 2015 AMIA Clinical Research Informatics Summit*, San Francisco, CA: USA, March 2015

Zhao L, Lim Choi Keung SN, Golby C, Ethier JF, Curcin V, Bastiaens H, Burgun A, Delaney BC, Arvanitis TN. "EU FP7 TRANSFoRm Project: Query Workbench for Participant Identification and Data Extraction", Papers/Podium Presentations. *In 2015 AMIA Clinical Research Informatics Summit*, San Francisco, CA: USA, March 2015

Lim Choi Keung SN, Zhao L, Curcin V, Ethier J-F, Burgun A, McGilchrist MM, Brodka P, Tuligowicz W, Delaney BC, Arvanitis TN, Andreasson A. Transform: Implementing a Learning Healthcare System in Europe through Embedding Clinical Research into Clinical Practice. *In: 2015 48th Hawaii International Conference on System Sciences (HICSS).* 2015. p. 3176–85.

Lim Choi Keung SN, Ethier J-F, Zhao L, Curcin V, Arvanitis TN. “The Integration Challenges in Bridging Patient Care and Clinical Research in a Learning Healthcare System” *ICEH 2014 at the European Medical Informatics Conference - MIE 2014* – Istanbul: Turkey August 2014

Curcin V, Arvanitis TN, Ethier J-F, Fraczkowsk K, Kazienko P, Brodka P, Andreasson A, Blizniuk G, Lim Choi Keung SN, Zhao L, Misiaszek A, McGilchrist M, Delaney B, Burgun A. “Semantic approach to achieving interoperability between clinical care and clinical research”. *In: New Routes for General Practice and Family Medicine - The 19th WONCA Europe Conference*. Lisbon: Portugal, July 2014

Lim Choi Keung SN, Zhao L, Rossiter J, McGilchrist MM, Culross F, Ethier J-F, Burgun A, Verheij R, Khan N, Taweel A, Curcin V, Delaney B, Arvanitis TN. “Detailed Clinical Modelling Approach to Data Extraction from Heterogeneous Data Sources for Clinical Research”. Papers/Podium Presentations. *In 2014 AMIA Clinical Research Informatics Summit*, San Francisco, CA: USA, April 2014

Lim Choi Keung SN, Zhao L, Rossiter J, Ogunsina I, Curcin V, Danger R, McGilchrist MM, Ethier J-F, Ohmann C, Kuchinke W, Taweel A, Delaney BC, Arvanitis TN, “Provenance-aware Query Formulation Tool to Identify Eligible Clinical Research Participants”, *In 2013 AMIA Clinical Research Informatics Summit*, San Francisco: USA, March 2013.

Delaney BC, Ethier J-F, Curcin V, Corrigan D, Friedman C. “International perspectives on the digital infrastructure for The Learning Healthcare System”. *AMIA 2013 Annual Symposium.* 19 November 2013. Washington, USA.

Delaney BC, Burgun A, Ethier J-F, Taweel A, Arvanitis T N. “Semantic Interoperability for Clinical Research in Europe: TRANSFoRm”, *In AMIA 2012 Summit on Clinical Research Informatics*, San Francisco, CA: USA, March 2012

Sommaire

Introduction

Translational research has been previously described using the acronym B2B: bench to the bedside. In this paradigm, research ideas emerge from fundamental research activities and the challenge identified for translational research was how transfer research outcome more effectively to the bedside in order to improve patient care. (Rodger 2000) The concept evolved over the following years to emphasise the bidirectional cycles from the "bench" to the "bedside" and back to the "bench": learning from the clinical studies outcomes and clinical care. This evolution was necessary as more genotype and phenotype data became available. (Burgun and Bodenreider 2008; *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease* 2011)

Today, multiple data points are created during healthcare delivery. While traditionally starting at the "bench", data mining has now changed the landscape and hypothesis can be first generated from care data. (Landman *et al.* 2010) Nonetheless, healthcare relies on complex ecosystems where patients interact with various institutions and care providers for punctual events (e.g. a pneumonia) or through multiple, interrelated visits across years of follow-up (e.g. diabetes or hypertension). (Galvin *et al.* 2015)

While initially concentrated in larger health care centres and hospitals, clinical data is now routinely captured through electronic means also during primary care activities. (Barker and Heisey-Grove 2015) Moreover, patients can participate in biomedical research protocols where data is also generated. (Lim Choi Keung *et al.* 2014) Finally, clinical support tools are becoming more prevalent and they also generate data points during their activities. (Musen *et al.* 2014) This creates significant challenges to derive knowledge from the patients' interactions throughout the healthcare system.

The task of linking patient data across multiple sources requires the development of standard methods for representing information, including controlled vocabularies and ontologies. (Burgun 2006) The re-use of data in the development of new hypotheses requires repositories such as the UMLS to enable conceptual linkage across coding systems. (Bodenreider 2004) Systems such as LexEVS have been developed that facilitate the management of lexical and semantic biomedical resources. (LexEVS) Nevertheless, multiple challenges remain.

Care delivery

In itself, care delivery presents many challenges in order to follow and understand the various processes at hand. Firstly, various clinical and administrative activities have different requirements and often use different informational systems to support them (e.g. laboratory tests, billing and accounting, operating room planning, clinical documentation, medication delivery). The underlying process and data models for each system will vary with each type of activity but also between the various implementations proposed by different vendors (e.g. the different electronic health record – EHR – "brand"). (Pecoraro *et al.* 2015) Secondly, depending on the type of care provided by each group of professionals or setting (e.g. primary care vs tertiary care hospitals), the scope of data used as well as its granularity will vary tremendously. (CPRD; Ontario Cancer Registry) The way it is collected and stored will also have a profound impact on its usability to answer a specific question. Thirdly, information sources will often be scattered geographically and administratively across various sites, institutions and even sometimes countries. As a result, patients' data is fragmented and obtaining a complete picture for a patient is quite challenging.

Research

Clinical research activities rely heavily on the participation of patients as subjects to permit knowledge discovery through retrospective and prospective research activities. This is essential to derive meaningful and pertinent knowledge to use as input for knowledge transfer activities. However, research is currently facing important difficulties as it is very resource-intensive (e.g. research assistants, data handling, regulatory activities…). It also faces recruitment challenges, especially with the advent of personalized medicine where treatments are targeted (and so tested) on subsets of disease patients (e.g. receptor-specific treatments in breast cancer). This is similarly true for research involving primary care where each site has a smaller number of patients and where patient characteristics will vary among practices, even within the same neighbourhood, making it challenging to select the optimal set of practices.

Knowledge transfer

Knowledge transfer encompasses multiple activities. It targets transfers at the population level through better, more adapted health policies but also at the patient level through guidelines and decision support. (Bernstein *et al.* 2015) As such, it presents significant challenges in context of translational research and clinical care. (Sarkar 2010) Firstly, clinical inertia, described as failure of health care providers to initiate or intensify therapy when indicated, impacts the quality of care provided to patients. Instituting change requires more effort than staying on a given course of action. (Phillips *et al.* 2001) It also contributes to increasing time between guidelines publications and their application in care settings. So despite knowledge generation from "the bench", translation into practice remains difficult. In order to address inertia, various informatics tools, generically identified as decision support systems (DSS), have been created to try to implement more effectively guidelines and other research outcomes into clinical practice. (Nazarenko *et al.* 2015) DSS are information management systems that involve knowledge representation, based on semantic structures such as ontologies, inferencing based on either rule-based or probabilistic approaches, and explanation, describing the decision making process. (Musen *et al.* 2014) Nevertheless, DSS are only effective when users deem the alerts to be relevant and therefore worth their time and attention. If alerts are too often inappropriate, they become ineffective and bothersome, leading to alert fatigue, a state in which the user becomes less responsive to alerts in general. This is especially prevalent when suggestions and alerts do not take into account specific patient or population characteristics like patients' comorbidities or disease prevalence in the practice's population.

Therefore, DSS in the context of translational medicine require access to computer-actionable, patient-level information to provide pertinent suggestions and should target situations and relevant to the population where they will be used. In order to achieve this, we need to better understand the target population to better orient research towards relevant questions and we must tailor the DSS operations to each individual patient. To realise this, we need data at a patient level and also at a population level.

After assessing challenges from care delivery, research and knowledge transfer (both at the patient level with DSS but also at the population level through health policies), it becomes apparent that being able to derive a full and clear picture from a patient's data is essential for translational research. (Sarkar 2010) Moreover, these activities cannot be considered as independent processes. Ensuring that they are addressed together in a coherent platform will address some of the challenges identified above.

Learning health systems

A possible answer started to emerge a few years ago as the Learning Health System (LHS). The pieces emerged gradually and the idea formalised by McGinnis and Friedman refers to the close coupling of practice of clinical medicine with both the conduct of research and the translation of research into practice as illustrated in Figure 1. (Westfall *et al.* 2007; Friedman *et al.* 2015)

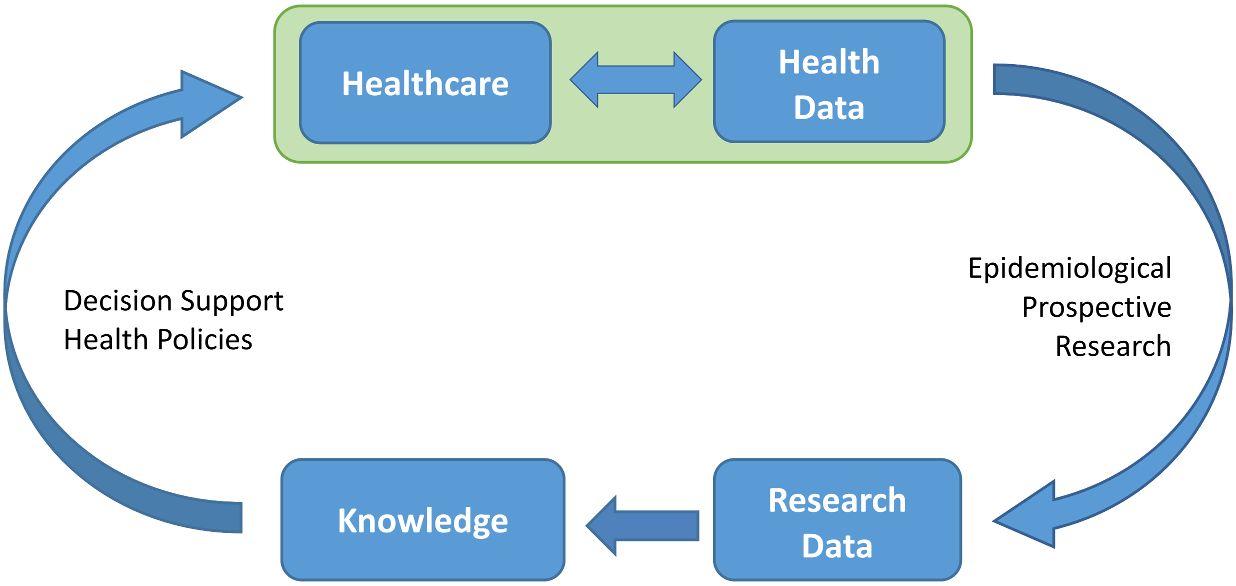

Healthcare
Health Data
Epidemiological Prospective Research
Research Data
Knowledge
Decision Support Health Policies

Figure 1: Learning Health System

This implies a shared semantics of clinical data exchanged between components of the system. With such a system in place, various sources from primary care to hospitals can be used to better understand patients, research can take place using real world patient data and knowledge transfer can be tailored to the population using a similar approach. DSS can also use the shared semantic data platform to contextualise alerts and optimise them for specific patient and situations, thereby increasing pertinence and so limiting alert fatigue.

At the core of the LHS lies a strong requirement to enable data flow with a shared semantics while integrating various clinical and research data sources. This presents significant challenges in terms of interoperability. The Translational Medicine and Patient Safety in Europe (TRANSFoRm) project is a research endeavour funded through the European Commission aiming at exploring requirements for a functional LHS in the

European context supporting primary care. (Delaney 2011) As such, it faces this challenge and the proposed approach has been deployed to support it.

This unified approach to support data requirements of the LHS will now be presented. Firstly, challenges and requirements will be described and existing approaches will be reviewed, including in the specific context of a LHS like TRANSFoRm, which is presented in an article published in the Biomedical Research International journal in 2015. (Delaney *et al.* 2015) This is the high-level article for TRANSFoRm. The co-authors of the article are in alphabetical order. Although not a work package leader, I have been invited as a co-author given my contribution to the project through the platform presented here.

Secondly, fundamental methodology for the new framework developed will be presented in an article published in the Journal of the American Medical Informatics Association in 2013. (Ethier *et al.* 2013) The first part of my contribution is presented as a definition of data requirements for a LHS. The novel idea of handling heterogeneity aspects through a unified platform, the evaluation of technical tools, the choice to use LexEVS and the development of the various model types (source, conceptual, mappings) to ensure sufficient expressivity and flexibility are original contributions from my work. Technical operationalisation through the data source connectors was developed by M. McGilchrist.

Thirdly, the implementation of the framework for the TRANSFoRm project is illustrated in an article published in the journal Methods of Information in Medicine ahead of print in 2014 in the focus theme issue “Managing Interoperability and Complexity in Health Systems”. (Ethier *et al.* 2015) The decision to use an ontology as the conceptual model (CDIM), its clinical content and its design choices are results of my work. Review of the content for its alignment with BFO and the realist paradigm has been achieved in collaboration with A. Barton.

Fourthly, the resulting platform and its contribution to the integration of prospective research and clinical care will be presented in a fourth article submitted to the Journal of the American Medical Informatics Association in 2016. The conceptual use of the platform models to achieve this goal through the linkage of CDIM with CDISC ODM (see article) is also an original contribution of my research. Obviously, various discussions with A. Burgun, B. Delaney and V. Curcin and M. McGilchrist have helped me to polish the approach and to integrate my original design with the TRANSFoRm project overall architecture.

Biomedical interoperability challenges in the context of the learning health system (LHS)

The learning health system concept is based on a close coupling between care, research and knowledge translation. These tight links are supported by intensive data exchange between the various components of the systems. Nevertheless, as described previously, a wide variety of systems must participate and support meaningful data exchange. This raises significant problems that must be overcome in order to enable proper data flow. One of the most significant ones is the heterogeneity between the various data sources to be used.

Moreover, it is now established that the study of complex diseases requires the effective integration of genomic, phenotypic and environmental data. The heterogeneity of the data generates significant challenges for integrating genomic information with relevant clinical data in a form that can be used to either test current hypotheses or generate new ones. (Sarkar *et al.* 2011)

Data sources heterogeneity

Heterogeneity stems from different characteristics of data sources. Firstly, the structure of the data source itself can be based on different technologies (e.g. relation databases, Extensible Markup Language – XML – documents, comma separated values, etc.). Moreover, each technology can allow a specific biomedical domain to be modelled differently based on the author of the model. Content elements will be grouped differently, names of tables and fields can be chosen arbitrarily. The level of granularity can also vary. In context of healthcare, while many types of informational elements will be common between the various systems (e.g. name, date of birth, patient record number), they will often be modelled differently based on the primary requirements of the system at hand. These common elements might be stored differently in a laboratory system and in a dialysis system for example. This aspect is identified in this document as structural heterogeneity.

Secondly, in order facilitate exchange of information, multiple controlled vocabularies and classifications have been created to support biomedical activities. (Rector 1999; Bodenreider 2004; Cimino 2011) They allow consistent use of shared terms in order to express some piece of knowledge. International examples range from the International Classification of Diseases (ICD) to the International Classification of Primary Care (ICPC), or Systematized Nomenclature of Medicine (SNOMED). (WHO | International Classification of Diseases (ICD); WONCA | International Classification of Primary Care; SNOMED CT) But then again some classifications are also used at a national level like Read Codes in the United Kingdom or even at a local level. (Read Codes) Moreover, many value sets (e.g. 1 for men and 2 for women) are used internally to code data in various health systems. These local codes are not standard and might be proprietary in nature. Overall, this aspect is identified as terminological heterogeneity in this document.[1]

Structural and terminological model interdependence

Although these two aspects can be described separately, they are strongly interdependent as has been demonstrated by A. Rector. (Qamar *et al.* 2007; Rector *et al.* 2009) In fact, to be able to derive the full semantics of a piece of information, both structural and terminological models need to be bound together.

For example, a terminological code like "E11" as found in the ICD version 10 does contain important information to understand the data at hand. (ICD-10 Version:2008 - E10) Its label is "Type 2 diabetes mellitus". It also provides inclusions and exclusions to further precise which pathologies should be represented by this code (e.g. exclusion of diabetes mellitus in pregnancy). Nevertheless, this only provides partial knowledge of the information at hand. Does it represent a diagnosis for the patient at hand (the patient has diabetes), or a diagnosis of a family member of the index patient (the patient's mother has diabetes)? Is it an admission diagnosis (the first hypothesis when the patient comes to seek care) or a final diagnosis provided after all relevant tests were obtained? Is it a diagnosis made by a medical student or an attending physician? For punctual events like pneumonias (J18 in ICD 10), does it represent that the patient currently has pneumonia or that he had a pneumonia in the past? (ICD-10 Version:2008 - J18) For episodic diseases like major depression (F32 in ICD 10), does a second code for a same patient represent a follow-up for a single episode of depression or does it represent a new episode? (Soler *et al.* 2012b; ICD-10 Version:2008 - F32) These issues are especially relevant when data from different institutions or domains needs to be used, for example when including data from primary care or from research platforms. (Soler

[1] Some authors use the term "semantic heterogeneity" to describe this but as demonstrated later in this work, semantics can only be fully derived when taking into account both structural and terminological aspects.

et al. 2012a) As a result, without knowing the semantic aspects carried by the structural (or information) model, it is impossible to answer these questions.

While this challenge has been identified a few years ago, it was at the time focusing mostly on clinical information. The LHS are now challenged to develop approaches that can bridge genomic, proteomic and even epigenomic information and place it into clinical context, adding a new layer of complexity. (Masys *et al.* 2012)

Other requirements

As multiple institutions from different components (care, research, and knowledge translation) are to participate in a LHS, each with different mandates and legal frameworks, creating a simple copy of all data into a central location is not possible. (Bastião Silva *et al.* 2015) Most of these institutions also currently have data sources structured in a certain way based on their current mission and also sometimes for historical reasons. (Friedman *et al.* 2015) The LHS does not control these organisations and cannot impose modifications to the existing data structures. It cannot either force EHR vendors to structure data differently. Given the variety of projects institutions must support, we also need to minimize resources required from an institution to participate in the LHS.

The LHS also needs to support prospective approaches. As opposed to other domains, many queries to be executed in the LHS are not known ahead of time. New concepts and new ways to interact with data will emerge and the LHS needs to be able to support them. Secondly, all sources are not identified on day one. The LHS needs to be designed to be organic in nature. It needs to be able to accept new sources as it grows and might also loose contact with others. Consequently, it cannot rely on content of sources available on day one and on query requirements elicited by a focus group of users to build a static approach to interoperability.

Review of approaches to address interoperability requirements of LHS including primary care.

In this section, approaches to data integration will be reviewed and analysed using requirements previously presented. The first two articles, presented in the methods section also address the issue and include specific examples of projects leveraging the different approaches but we will summarize the principles here. Please see the articles themselves for more specific examples of projects.

Approaches to data integration can be broadly classified along two categories: Data warehousing (DW) and data mediation (DM). (Lenzerini 2002; Hernandez and Kambhampati 2004; Louie *et al.* 2007) Also encountered in the literature, data federation (DF) can be seen as a specific case of DM where each source uses the same structure.

Data warehousing is possibly the most notorious approach with initiatives like Informatics for Integrating Biology & the Bedside (I2B2) in the biomedical domain. (Mansmann *et al.* 2009; Murphy *et al.* 2010: 2) DW is a common approach to integrate various data sources within an institution. (Huser and Cimino 2013) Using the Extract-Transform-Load (ETL) approach, data from each individual source is extracted, transformed to fit the structure of the data warehouse and then load into it. (Vassiliadis *et al.* 2002) In recent years, multidimensional structures (e.g. star schema) have been used but other modelling approaches are also possible. (Abelló *et al.* 2001) For more details on data warehousing approaches and their characteristics, one can refer to a recent review. (Khnaisser *et al.* 2015) In the context of a LHS it becomes readily apparent that it cannot become the central platform for data exchange. It requires data leaving institutional boundaries to be entirely stored in a single central repository. As mentioned previously, this is not possible in our use case.

Data federation can present a very interesting profile for data integration. DF is a network of data sources structured identically and distributed among different sites and often different institutions. The electronic Primary Care Research Network (ePCRN) in the primary care domain is a well-known exemplar of data federation. Its goal is to enable the development of an electronic infrastructure to support clinical research activities in primary care practice-based research networks. It has been developed with sites both in the United

States and in the United Kingdom. Data in ePCRN must be structured according to the American Society for Testing and Materials continuity of care record information model. (Peterson *et al.* 2006; Delaney *et al.* 2012) Similarly, SHRINE is an initiative that federates sites using I2B2 as data repositories. (Weber *et al.* 2009)

Data federation allows data to reside in each institution and be transmitted only when needed and when allowed, for specific reasons. Since everyone shares the same data modelling, the same query can be run at each site and data easily aggregated. Achieving this requires coordination between sites and agreements that each site will use the same data structure. In context of a LHS going from primary care to quaternary specialized care, this is not possible. Moreover, multiple institutions already have data repositories or EHRs. The LHS has no way to enforce a change at participating institutions to adopt a shared data structure or terminology.

While, DF is a specific form of data mediation, more generic approaches can also be identified in the literature for DM. (Wiederhold 1992) In its more generic form, it uses a central model (also identified as a conceptual mediation model) designed to support query expression sent to the system. (Ashish *et al.* 2010) Local models are also produced to represent the content of the data sources (structural models). When using it, queries issued based on the central model are then translated locally for each source in order to create a relevant query that can be executed on the data source directly. Data is then returned centrally.

DM requires that the central model be mapped to each local model. To achieve this, one needs to be a “view” on the other. Following this, two sub-types of DM can be defined: global-as-view (GAV) and local-as-view (LAV). (Calì *et al.* 2001) In GAV, the central model is derived by merging each local models according to pre-specified transformations. It is a direct reflections of available sources at a specific time. Even though historically it presented better performances, it also exposes a very dynamic model to the users of the system. In a situation where the responsible organisation for the central model do not control the sources, asynchrony and incoherence can occur when local sources change but the central model is not updated. As a result, a GAV approach would not fit the requirements for a LHS.

The local-as-view approach on the other hand derives its central model from the users’ requirements. It is built irrespectively of what data is available but rather based on what type of data would be useful. The local source models are then built to represent the data structure and are subsequently mapped to the central model. They are in effect views on the central model. Any data not mapped to the central model will not be available nor visible to the DM users. The resulting system provides a stable and relevant model to its users to support query and provides flexibility to map sources from various part of the target domain. It is therefore the approach chosen to support interoperability in the LHS.

While the local-as-view approach was a good candidate to support a LHS, it had not previously been deployed in this specific context. Even though caBIG and Advancing Clinico-Genomic Trials implemented this approach in the cancer research domain, it has not been used previously in primary care. (Stanford and Mikula 2008; Martin *et al.* 2011) Current approaches also treat structural and terminological models for mediation separately. We therefore propose here a unified interoperability approach to support interoperability in a LHS including primary care based on a DM LAV vision.

Implementation of a LHS in primary care: the TRANSFoRm project

The proposed approach was developed and implemented as part of the TRANSFoRm project (http://www.transformproject.eu/), a FP7 initiative funded via the Patient Safety Stream of ICT for Health. TRANSFoRm aims at implementing a LHS for primary care to foster patient safety and facilitate research in primary care.

In order to orient its development, three use-cases were designed. (Delaney *et al.* 2015) They cover the full LHS cycle. The first one mandates a retrospective research study using routine primary care data and genomic profiles from a biobank. The goal is to evaluate response variation of a specific diabetes medication class (sulfonylurea) depending varying genomic profiles. The study had been previously done in Scotland and its objective is to evaluate if a LHS could support design and execution of a similar study.

The second use-case aims focuses on a randomized control trial (RCT) comparing the use of anti-acid medication (proton pomp inhibitors) regularly or only as needed in patients with heart burns (gastro-oesophageal reflux disease) seen in primary care practices. Given the relatively small number of patients per practice (as opposed to hospitals), recruitment can be quite difficult and costly. (Mastellos *et al.* 2015) In order to evaluate the effect of a LHS on the conduct of RCT in primary care, the practices participating in the trial are also randomized. Some use "current standard" for patient recruitment (with research assistant manually screening patients) and the others use automated tools provided by the TRANSFoRm platform to facilitate recruitment and maximise automated use of existing data from the practice EHR to fill the research forms.

The last use-case focuses on knowledge transfer through a decision support tool. (Corrigan 2015) Using data from the EHR, the ontological approach classifies clinical cues and analyses them in order to provide highly relevant alerts and suggestions to primary care clinicians trying to establish a diagnosis for three common presenting symptoms: chest pain, abdominal pain and shortness of breath.

The following article from Biomedical Research International published in June 2015 presents in more details the project and its various components. (Delaney *et al.* 2015)

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 961526, 8 pages
http://dx.doi.org/10.1155/2015/961526

Research Article

Translational Medicine and Patient Safety in Europe: TRANSFoRm—Architecture for the Learning Health System in Europe

Brendan C. Delaney,[1] Vasa Curcin,[1] Anna Andreasson,[2] Theodoros N. Arvanitis,[3] Hilde Bastiaens,[4] Derek Corrigan,[5] Jean-Francois Ethier,[6] Olga Kostopoulou,[1] Wolfgang Kuchinke,[7] Mark McGilchrist,[8] Paul van Royen,[4] and Peter Wagner[9]

[1]*King's College London, London SE1 3QD, UK*
[2]*Karolinska Institutet, 14183 Stockholm, Sweden*
[3]*University of Warwick, Coventry CV4 7AL, UK*
[4]*University of Antwerp, 2610 Antwerp, Belgium*
[5]*Royal College of Surgeons of Ireland, Dublin 2, Ireland*
[6]*INSERM, 6 Paris, France*
[7]*University of Düsseldorf, 40225 Düsseldorf, Germany*
[8]*University of Dundee, Dundee DD2 4BF, UK*
[9]*Quintiles GmbH, 63263 Neu-isenberg, Germany*

Correspondence should be addressed to Brendan C. Delaney; brendan.delaney@kcl.ac.uk

Received 16 February 2015; Accepted 8 June 2015

Academic Editor: Danielle Dupont

The Learning Health System (LHS) describes linking routine healthcare systems directly with both research translation and knowledge translation as an extension of the evidence-based medicine paradigm, taking advantage of the ubiquitous use of electronic health record (EHR) systems. TRANSFoRm is an EU FP7 project that seeks to develop an infrastructure for the LHS in European primary care. *Methods*. The project is based on three clinical use cases, a genotype-phenotype study in diabetes, a randomised controlled trial with gastroesophageal reflux disease, and a diagnostic decision support system for chest pain, abdominal pain, and shortness of breath. *Results*. Four models were developed (clinical research, clinical data, provenance, and diagnosis) that form the basis of the projects approach to interoperability. These models are maintained as ontologies with binding of terms to define precise data elements. CDISC ODM and SDM standards are extended using an archetype approach to enable a two-level model of individual data elements, representing both research content and clinical content. Separate configurations of the TRANSFoRm tools serve each use case. *Conclusions*. The project has been successful in using ontologies and archetypes to develop a highly flexible solution to the problem of heterogeneity of data sources presented by the LHS.

1. Introduction

The Learning Health System (LHS) describes an approach to improve healthcare that is solidly founded on the creation and use of knowledge; "health" as opposed to "healthcare" is sometimes used to emphasise the role of consumers as cocreators and users of health knowledge [1]. The development of the LHS is a natural outcome of the evolution of evidence-based medicine (EBM). Based on the greater utilisation of electronic health records (EHRs) and on novel computing paradigms for data analysis, the LHS provides potential solutions for the glacial slowness of both the traditional research process and the research translation into improved care [2].

EBM is focused on generating medical evidence and using it to make clinical decisions. The highest level of evidence, level 1 evidence of the effectiveness of a healthcare intervention in EBM, consists of a meta-analysis of randomised controlled trials (RCTs) [3]. However, RCTs are complex and extremely expensive, the result being that much of healthcare remains unsupported by high quality evidence. Furthermore, RCTs themselves are prone to bias and manipulation in the choice of eligible subjects, comparators, and outcome measures [4]. One solution has been to carry out light touch and simple, termed "pragmatic" RCTs with very inclusive eligibility criteria and followup via routine data collection. It is those kinds of RCTs that lend themselves most to incorporation into a LHS.

There is also potential to replace RCTs with analysis of routine data, using techniques such as instrumental variables and propensity scores to control for bias [5]. Much future research is needed to define when routine data could be a sufficient answer to a problem and when an RCT is required. Furthermore, healthcare practice is not solely limited to interventions, but diagnosis and prognostication play essential parts and are underpinned by prospective cohort evidence. Again, routine data could play a significant role in replacing time-consuming and costly cohort designs.

Primary healthcare is the first point of contact with health services of patients with undifferentiated problems and also provides continuing care for patients with chronic diseases and follows families from "cradle to grave." These functions present a particular problem for EBM. The vast majority of research, be it diagnostic or intervention based, takes place in specialist centres and in highly selected populations [6]. Diagnostic features are not portable across populations with different prevalence and spectrum of disease. Likewise, patients in RCTs are younger and fitter, take fewer drugs concurrently, and have less comorbidity than typical primary care populations. Therefore, many RCTs suffer from limited external validity [7].

Even if appropriate research evidence exists, it is unlikely to be available at the point of care. Early formulations of EBM typically applied to the highly motivated clinician who formulates questions during clinical practice and searches for evidence. Indeed, Professor Sackett's team at Oxford developed an "evidence cart" for ward rounds, with a copy for MEDLINE and a projector to assist in this process in real time [8]. Over the subsequent years, the process of knowledge translation has become formalised: guidelines are explicitly built on systematic reviews of the best available evidence and are refined down to a series of statements to support clinical care, with an associated level of supporting evidence and strength of recommendation [9]. However, even in countries like the UK, where a national agency (National Institute for Health and Care Excellence) is funded to carry out this process, guidelines may only be updated once in a decade. Increasingly, the number of potential guidelines applicable to a given patient at a given point on the care pathway becomes a problem of memory and prioritisation for the clinician, let alone the patient. The LHS offers a potential means of using highly advanced electronic triggers to help with advising when one treatment or diagnosis is favoured. It should also be possible to reintroduce patient choice by explicit weighting of options using patient-derived outcome data.

The LHS concept is still in its infancy, and much needs to be done to explore and demonstrate the potential for using an advanced digital infrastructure to support the LHS. The FP7 TRANSFoRm project (http://www.transformproject.eu/) was funded via the Patient Safety Stream of ICT for Health. Efficient research design and knowledge translation are a core underpinning of safe clinical practice. It is not good enough to simply avoid error, defined as care that falls well below the average standard, but clinicians should be seeking optimal care for their patients. The LHS, at its barest essential, is all about promoting optimal care. The TRANSFoRm project aimed to develop and demonstrate methods, models, standards, and a digital infrastructure for three specific components of the LHS:

(1) genotype-phenotype epidemiological studies using multiple existing primary care and "biobank" genomic datasets;

(2) RCTs with both data and trial processes embedded within the functionality of EHRs and the ability to collect Patient Reported Outcome Measures (PROMs) on demand;

(3) decision support for diagnosis, based on clinical prediction rules (best diagnostic evidence) and fully integrated with a demonstrator EHR system.

2. Methods

Each specific clinical "use case" (shown below) served four purposes: initial requirements elicitation; detailed modelling of infrastructure and required data elements; design of concurrent validation and evaluation studies; and final clinical demonstrations. 21 partner organisations in ten EU member states took part in the project, over five years. At the time of writing, the project has 11 months to run and the final evaluation and clinical studies are about to commence.

TRANSFoRm Use Cases

Diabetes Use Case. The aim of the Diabetes use case is to enable a distributed query to look for eligible patients and extract data from multiple federated databases. In the pilot study, the query will define patients and data to support analysis of the relationship between well-selected single nucleotide polymorphisms (SNPs) in type 2 diabetic patients and the response to sulfonylurea.

GORD Use Case. The aim of the GORD use case is to investigate the effectiveness of on demand versus continuous use of proton pump inhibitors on reflux symptoms, quality of life, and self-rated health in patients with gastrooesophageal reflux disease in primary care. The study will be conducted in five localities (UK: two vendors, Poland, Netherlands, and Crete) and it will aim to recruit, randomise, and follow 700 patients at 40 primary care centres using the clinical trial application.

Diagnosis Use Case. The aim of the diagnosis use case is to provide integrated point-of-care decision support for patients presenting with chest pain, abdominal pain, and shortness of breath.

TRANSFoRm aims to produce a highly flexible infrastructure that presents the lowest possible barriers to entry for EHR systems and datasets, but at the same time it makes the maximum use of the existing data standards and methods for managing heterogeneity, both structural and terminological, between data sources. A basic principle of the TRANSFoRm project was to use available standards and models as much as possible and integrate them into the TRANSFoRm infrastructure. It was decided early on in the project that TRANSFoRm would take a model-based approach, using 4 models to capture (1) clinical meaning, (2) research meaning, (3) provenance, and (4) diagnostic meaning. The latter is essentially a subset of the clinical model, but it was modelled separately for efficiency. The archetype approach of constraining one model against the other, in a two-level design (clinical and research), was used to describe data elements [10]. Where available, existing tools for building and maintaining models as an ontology were used, although we presented a novel use of LexEVS, which we employed to support both structural and semantic models [11].

Clinical concepts were modelled using an ontology (termed the Clinical Data Information Model, CDIM) [12]. Additional semantic detail for data elements was expressed by using LexEVS to support binding of terminology terms to CDIM expressions. For representation of research processes, we extended an existing domain model, the Primary Care Research Object Model, adding objects primarily in the clinical area [13]. The resulting Clinical Research Information Model (CRIM), in conjunction with CDIM, enabled a two-level archetype to be defined for each required data element in the use cases. In order to define case report forms and study designs for the RCT, we used the CDISC ODM and SDM standards, but adding an archetype approach for the description of the data element "payload" [14].

The intention from the outset with TRANSFoRm was that all models would be published, standards would be reused and adapted as required, the software would reuse the existing open source components, if available, and all TRANSFoRm software components would be made available as open source tools under an "Apache" license. We believe that the value lies in the data and the knowledge generated from it and that amortizing the infrastructure can only act as a potential barrier for realising the value of the data/knowledge.

Evaluation of TRANSFoRm will consist of a technical validation of the TRANSFoRm tools and three clinical and sociotechnical evaluation studies. For the DSS, an evaluation of the system, integrated with the In Practice Systems Vision 3 EHR system, is underway. General practitioners are conducting a simulated clinical session with actors simulating patients presenting with carefully prepared test problems. This is a within-subjects design, with the cases solved first without and then with the DSS and the primary outcome being accuracy. We also measure usability and amount of information coded into the EHR. The Diabetes use case is being evaluated on the basis of performance, as judged by users, of the system in selecting and extracting data from five databases. Accuracy of selecting eligible patients by users employing the TRANSFoRm Query Workbench will be measured. The GORD (gastrooesophageal reflux disease, a disorder caused by the retrograde flow of gastric contents from the stomach into the oesophagus, causing symptoms and/or mucosal damage) study is being conducted as a full clinical RCT (individual subjects randomised) with a nested evaluation study. Principal outcomes of the clinical study are symptom profiles and quality of life measured by PROMs (Patient Reported Outcome Measures) collected on smartphones via a dedicated TRANSFoRm mobile data collection app. The sociotechnical evaluation is a nested cluster trial and will compare recruitment rates, completeness of data, and costs of the TRANSFoRm system compared to usual practice, in this case, a simple web form for the clinical measures and paper questionnaires for the PROMs. The results of the three TRANSFoRm evaluation studies will be available in late 2015.

3. Results

The TRANSFoRm software ecosystem is comprised of a set of generic middleware components that provide essential shared functions for the LHS applications built in TRANSFoRm, namely, secure data transport, authentication, semantic mediation, and data provenance (with respect to processing of data within TRANSFoRm). As LHS is characterized by routine production, transformation, and dissemination of data and knowledge, secure channels and reliable authentication are necessary to ensure confidence and buy-in by the data owners. The data itself resides in a vast array of distributed repositories that vary both in structure and in terminology, making data interoperability a key requirement that TRANSFoRm delivers using a semantic mediation approach combined with the standard data connectivity module (data node connector: DNC). The DNC implements data interoperability, as well as managing workflow processes and data extraction for participating EHRs and data sources, as discussed in the next section. Different flavours of DNC operate in epidemiology and RCT use cases, as the RCT DNC has to support additional requirements of the RCT workflow. Data provenance capture in TRANSFoRm implements traceability, which is necessary both to support trust and transparency and to enable learning and improvement in LHS processes.

On top of these shared components, three application specific tools were built to support the use cases: epidemiological study query workbench, clinical trial monitoring tool, and a diagnostic support plugin for EHR systems.

The high-level overview of the software components is shown in Figure 1.

4. Epidemiological Study Application

The epidemiological study TRANSFoRm software configuration (Figure 2) is used in the genotypic-phenotypic T2D study use case and consists of tools for secure, provenance-enabled design and execution of eligibility queries and data extractions from heterogeneous data sources. Eligibility queries are

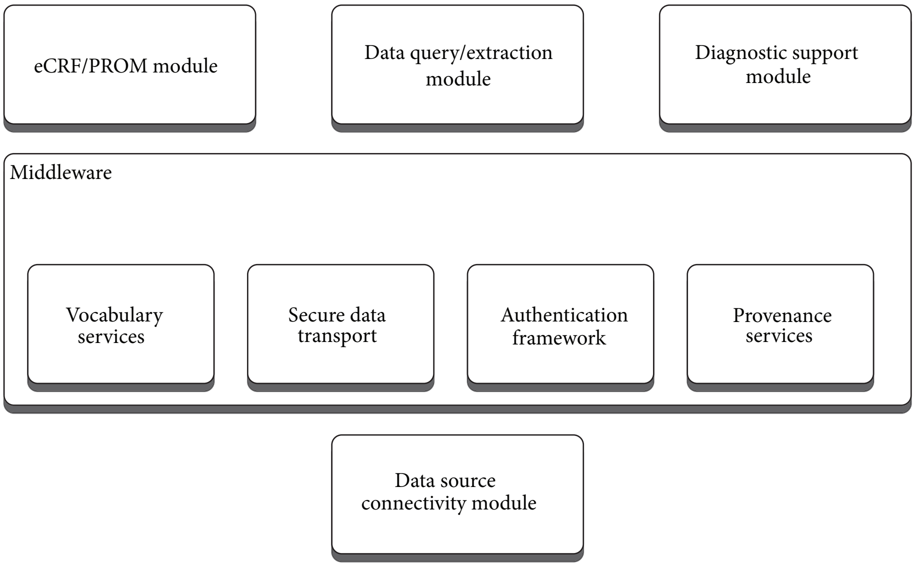


Figure 1: High-level software components.

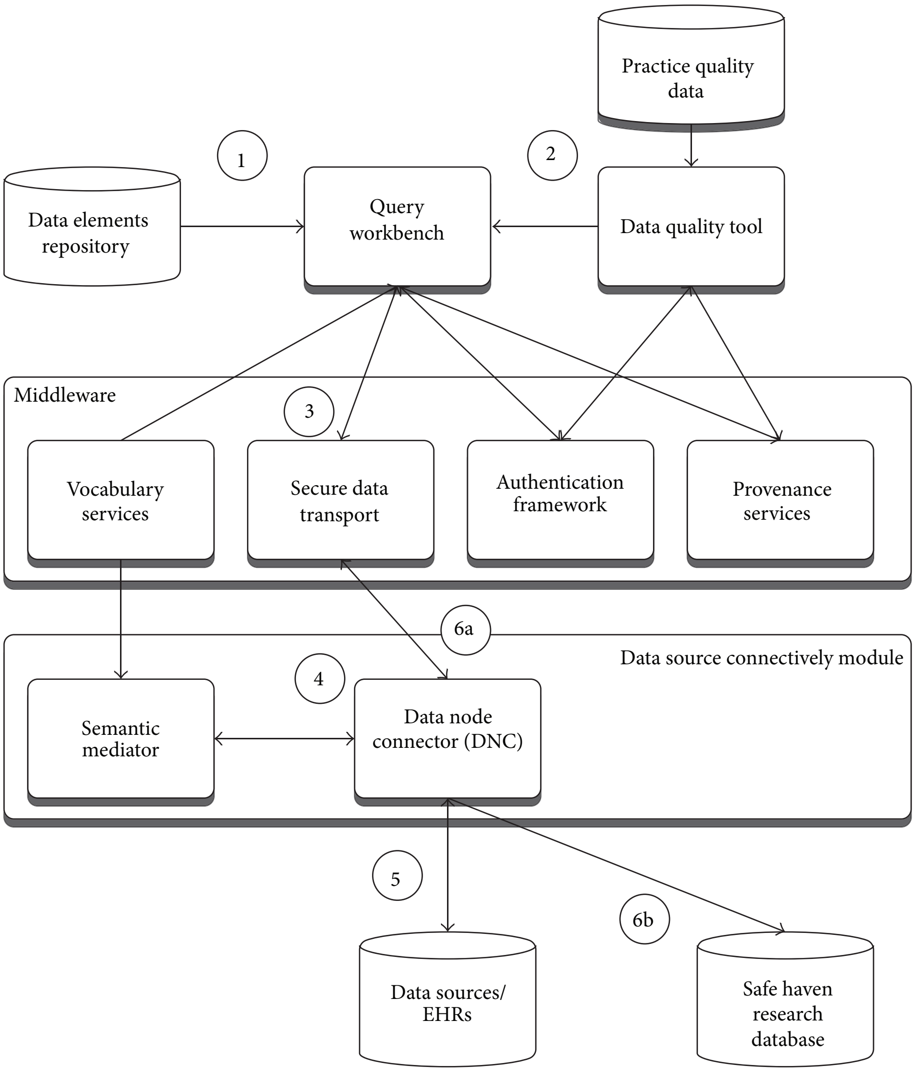
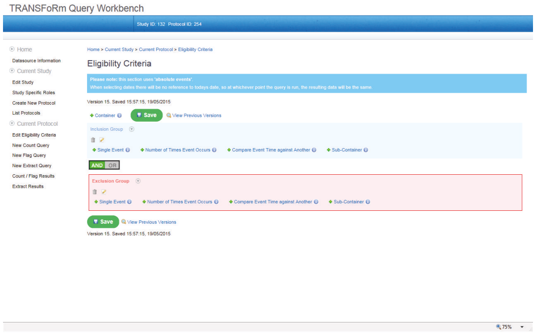


Figure 2: Epidemiological study configuration annotated with steps in the query process.



TRANSFoRm Query Workbench

Study ID: 132 Protocol ID: 254

- Home
- Datasource Information
- Current Study
- Edit Study
- Study Specific Roles
- Create New Protocol
- List Protocols
- Current Protocol
- Edit Eligibility Criteria
- New Count Query
- New Flag Query
- New Extract Query
- Count / Flag Results
- Extract Results

Home > Current Study > Current Protocol > Eligibility Criteria

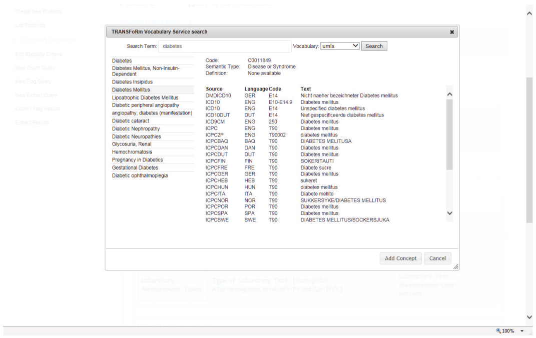
Eligibility Criteria

Please note: this section uses 'absolute events'.
When selecting dates there will be no reference to todays date, so at whichever point the query is run, the resulting data will be the same.

Version 15. Saved 15:57:15, 19/05/2015

Version 15. Saved 15:57:15, 19/05/2015

Figure 3: TRANSFoRm Query Workbench.



TRANSFoRm Vocabulary Service search

Search Term: diabetes Vocabulary: umls

- Diabetes
- Diabetes Mellitus, Non-Insulin-Dependent
- Diabetes Insipidus
- Diabetes Mellitus
- Lipoatrophic Diabetes Mellitus
- Diabetic peripheral angiopathy
- angiopathy; diabetes (manifestation)
- Diabetic cataract
- Diabetic Nephropathy
- Diabetic Neuropathies
- Glycosuria, Renal
- Hemochromatosis
- Pregnancy in Diabetics
- Gestational Diabetes
- Diabetic ophthalmoplegia

Code: C0011849
Semantic Type: Disease or Syndrome
Definition: None available

Source	Language	Code	Text
DMDICD10	GER	E14	Nicht naeher bezeichneter Diabetes mellitus
ICD10	ENG	E10-E14.9	Diabetes mellitus
ICD10	ENG	E14	Unspecified diabetes mellitus
ICD10DUT	DUT	E14	Niet gespecificeerde diabetes mellitus
ICD9CM	ENG	250	Diabetes mellitus
ICPC	ENG	T90	Diabetes mellitus
ICPC2P	ENG	T90002	diabetes mellitus
ICPCBAQ	BAQ	T90	DIABETES MELITUSA
ICPCDAN	DAN	T90	Diabetes mellitus
ICPCDUT	DUT	T90	Diabetes mellitus
ICPCFIN	FIN	T90	SOKERITAUTI
ICPCFRE	FRE	T90	Diabete sucre
ICPCGER	GER	T90	Diabetes mellitus
ICPCHEB	HEB	T90	sukeret
ICPCHUN	HUN	T90	diabetes mellitus
ICPCITA	ITA	T90	Diabete mellito
ICPCNOR	NOR	T90	SUKKERSYKE/DIABETES MELLITUS
ICPCPOR	POR	T90	Diabetes mellitus
ICPCSPA	SPA	T90	Diabetes mellitus
ICPCSWE	SWE	T90	DIABETES MELLITUS/SOCKERSJUKA

Figure 4: Concept search in TRANSFoRm Query Workbench.

formulated by the researcher in the query workbench (QWB) web tool (Figure 3) using model-based constructs (Figure 2, step 1). QWB users enter clinical terms into the system which then presents the user with a list of corresponding concepts from standard terminologies and classifications (Figure 4). The researchers are able to use a data quality tool, storing metadata about available practices and data that reside in them, to restrict the search to practices with a high registration percentage of the variables targeted in the study (step 2). The queries are dispatched to the data sources via the middleware (step 3) to the local data node connector. This is a TRANSFoRm component that sits at the data source and translates the generic CDIM-based query into a local representation using the semantic mediator component (step 4) and subsequently presents that locally interpretable query either to the data source directly or to a human agent for final approval (step 5), before returning the result. Three types of queries are supported: patient counts, flagging patients, and data extraction. Results of count and flag queries are sent back to the query workbench via the middleware (step 6a) and can be viewed by the researcher in the QWB web tool. The patient data extraction result is passed to a safe haven (step 6b), accessible only to the authorised researcher, using the appropriate secure data transport mechanism.

5. Clinical Trial Application

The clinical trial software configuration (Figure 5) is used in the GORD use case and consists of components needed for design, deployment, and collection of trial data, backed by provenance and secure authentication framework for researchers. The trial data collection is supported using electronic Case Report Forms (eCRFs) and Patient Reported Outcome Measures (PROMs). The former are filled in via a web browser by the clinician, while the latter are completed by the patients using either web or mobile devices. Also supported is the orchestration of data collection across multiple clinical sites where the trials are taking place.

The TRANSFoRm architecture delivers important components of clinical trials: patient eligibility checks and enrolment, prepopulation of eCRF data from EHRs, PROM data collection from patients, and storing of a copy of study data in the EHR. The key component of the architecture is the TRANSFoRm Study System (TSS) that coordinates study events and data collections, using HTML form templates with bound queries for preloading data from the EHR. The studies, represented using a custom extension of CDISC SDM/ODM standard, are loaded into the TRANSFoRm Study System (step 1). Whenever an interaction is required between the Study System and EHR, for example, eligibility checks or partial filling of eCRF forms form EHR data, a query is fired off to the EHR via the data node connector (step 2). As in the epidemiological study configuration, the DNC acts as a single point of contact of TRANSFoRm components and the local EHR. In addition to translating and sending queries to the EHR (step 3), the DNC acts as a web server that displays eCRF forms for the clinician to fill with study-required information not present in the EHR. Once completed, the form is submitted to both the study database and the EHR for storage considering requirements for eSource data use in clinical trials (step 4). The message protocol for this interaction is currently undergoing comparison evaluation with the IHE standards [17]. The PROM data is collected directly from the patients using web or mobile devices (step 5). The software configuration for the GORD study undergoes a formal Computer System Validation (CSV) process including qualifications for installation, operation, and performance to ensure that study system and study process have been Good Clinical Practice- (GCP-) validated prior to being employed in the GORD clinical trial use case. Because of the narrow connection between EHR and study system, part of GCP-validation is the assurance of data privacy and confidentiality of the personal patient data.

6. Diagnostic Support Application

Diagnostic support software configuration (Figure 6: diagnostic support configuration) consists of tools for mining new rules from health data sources and managing their deployment into the knowledge base, upon which an evidence service is operating to drive a diagnostic support tool embedded into a local EHR system.

The primary function of the tool is to suggest to clinicians diagnoses to consider at the start of the clinical encounter based only on the existing information in the patient record and the current reason for encounter [18]. It also allows bottom-up input of observed patient cues (symptoms and signs), independent of associated diagnosis, or top-down

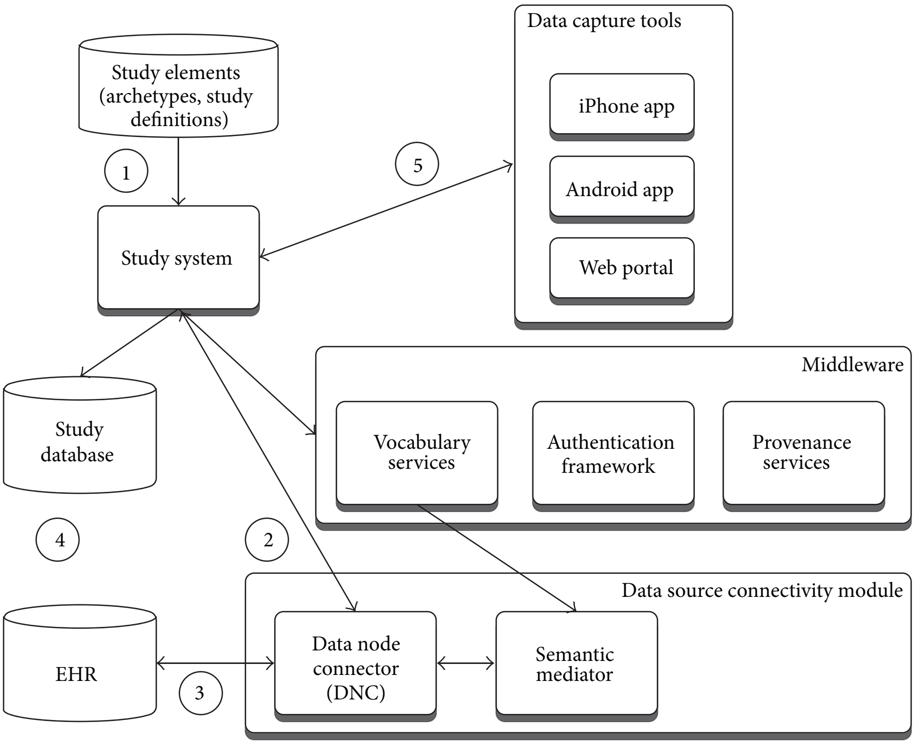


Figure 5: TRANSFoRm clinical trial configuration.

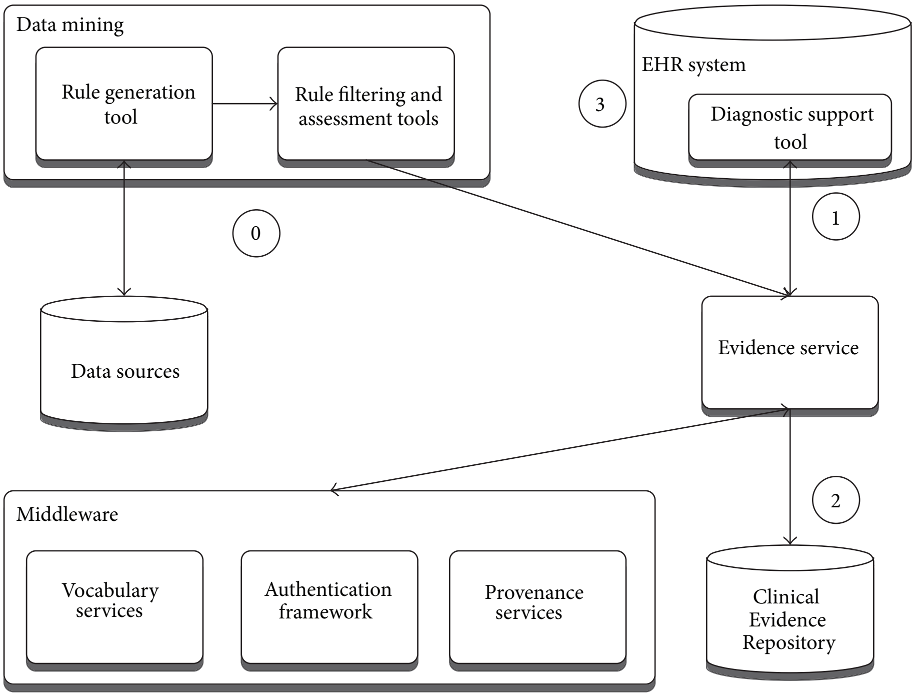


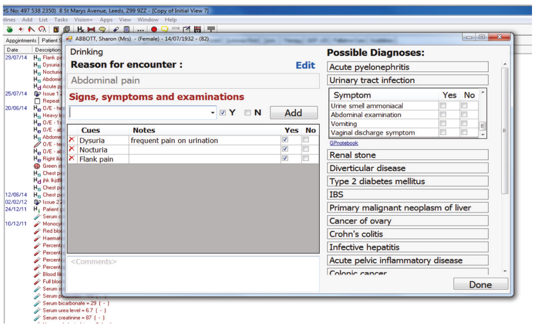
Figure 6: Diagnostic support configuration.

drilling into and selection of cues supporting specific diagnoses.

The rules used in the diagnostic process are generated by data mining tasks (step 0), which get manually curated and fed through the evidence service into the Clinical Evidence Repository. When the patient presents, the cues entered or selected are then used to dynamically rank the potential differential diagnoses (Figure 7). This is done by the DSS plugin embedded into the EHR, sending data to the evidence service (step 1), which queries the rules stored in

Table 1: A table of outputs and exploitation plans.

TRANSFoRm output	Exploitation plan
(1) Privacy model: a "zone" model with an explicit method of graphically depicting the zones and operation of filters between zones	Published method [15]
(2) Provenance infrastructure: based on the Open Provenance Model [REF], each infrastructure component captures a provenance trace that enables reconstruction of an audit trail for any given data element	Published method [16]
(3) Clinical prediction rule ontology based web service	The diagnostic ontology has been made available as a public download in OWL format on the TRANSFoRm website (http://www.transformproject.eu/). A future project is required to extend the data beyond the three initial reasons for encounter
(4) Research data model	CDIM [12] and CRIM [13] have been published. A full description of the use of CDIM and CRIM in the construction of data node connectors will be published and made available on the TRANSFoRm website
(5) eCRF	Extension of CDISC ODM and SDM by the incorporation of archetypes with references to the CRIM and CDIM models will be published and discussions are ongoing with CDISC regarding future incorporation into the standards. A reference implementation of the clinical trial system will be maintained within the European Institute. At present, individual archetypes have to be written by hand; discussions are in hand for the production of an archetype authoring tool
(6) Data federation	A reference implementation of the genotype-phenotype study system will be maintained within the European Institute. Search authoring tools will be available open source
(7) DSS integration	The DSS is currently integrated with the InPS Vision 3 system. Further work is required to move this to a data node connector/CDIM-based flexible system



HS No: 497 538 2350) 8 St Marys Avenue, Leeds, Z99 9ZZ - [Copy of Initial View 7]

ABBOTT, Sharon (Mrs) - (Female) - 14/07/1932 - (82)

Drinking

Reason for encounter : Edit

Abdominal pain

Signs, symptoms and examinations

☑ Y □ N Add

Cues	Notes	Yes	No
Dysuria	frequent pain on urination	☑	□
Nocturia		☑	□
Flank pain		☑	□

<Comments>

Possible Diagnoses:

Acute pyelonephritis

Urinary tract infection

Symptom	Yes	No
Urine smell ammoniacal	□	□
Abdominal examination	□	□
Vomiting	□	□
Vaginal discharge symptom	□	□

GPnotebook

Renal stone

Diverticular disease

Type 2 diabetes mellitus

IBS

Primary malignant neoplasm of liver

Cancer of ovary

Crohn's colitis

Infective hepatitis

Acute pelvic inflammatory disease

Colonic cancer

Done

Serum bicarbonate = 29 (-)

Serum urea level = 6.7 (-)

Serum creatinine = 87 (-)

Figure 7: Diagnostic support tool implemented as a plugin to InPS Vision EHR system.

the Clinical Evidence Repository (step 2), before sending the potential diagnoses back, annotated with levels of support and confidence for the presenting case. Upon exiting the tool, the coded evidence cues and current working diagnosis can be saved back to the patient EHR (step 3).

7. Conclusions

TRANSFoRm demonstrated how a Learning Health System can be implemented in European clinical research and practice. The full list of project outputs and the exploitation plan for each are shown in Table 1 and promoted via an open source model. TRANSFoRm will be a full participant in the European Institute for Innovation through Health Data and will make its tools and models available via the institute. In addition, we are internationally active as participants and promoters of the Learning Healthcare System. Via the LHS, we are publishing models, standards, and tools to the world research community. The UK serves as an exemplar of our business model, with multiple EHRs participating in the project as well as the Medicines and Healthcare Products Regulatory Agency, Clinical Practice Research Datalink (CPRD). CPRD currently extracts data from practices to a total population of 8 million and links them to 20 other health datasets. CPRD will be using the TRANSFoRm clinical trial tools, in conjunction with additional reworking by a commercial software vendor to create a full EHR-embedded clinical trial facility for the UK Clinical Research Network.

Disclaimer

The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under Grant Agreement no. 247787. Additional support was received 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.

References

[1] C. Friedman, J. Rubin, J. Brown et al., "Toward a science of learning systems: a research agenda for the high-functioning Learning Health System," *Journal of the American Medical Informatics Association*, vol. 22, no. 1, pp. 43–50, 2015.

[2] C. P. Friedman, A. K. Wong, and D. Blumenthal, "Achieving a nationwide learning health system," *Science Translational Medicine*, vol. 2, no. 25, Article ID 57cm29, 2010.

[3] D. L. Sackett, W. M. Rosenberg, J. A. Gray, R. B. Haynes, and W. S. Richardson, "Evidence based medicine: what it is and what it isn't," *British Medical Journal*, vol. 312, no. 7023, pp. 71–72, 1996.

[4] T.-P. Van Staa, B. Goldacre, M. Gulliford et al., "Pragmatic randomised trials using routine electronic health records: putting them to the test," *British Medical Journal*, vol. 344, no. 7843, article e55, 2012.

[5] R. L. Tannen, M. G. Weiner, and S. M. Marcus, "Simulation of the Syst-Eur randomized control trial using a primary care electronic medical record was feasible," *Journal of Clinical Epidemiology*, vol. 59, no. 3, pp. 254–264, 2006.

[6] B. Delaney, "Primary care research in the postmodern world," *Family Practice*, vol. 21, no. 2, pp. 123–124, 2004.

[7] M. Fortin, J. Dionne, G. Pinho, J. Gignac, J. Almirall, and L. Lapointe, "Randomized controlled trials: do they have external validity for patients with multiple comorbidities?" *The Annals of Family Medicine*, vol. 4, no. 2, pp. 104–108, 2015.

[8] D. L. Sackett, "Finding and applying evidence during clinical rounds—the 'evidence cart,'" *The Journal of the American Medical Association*, vol. 280, no. 15, pp. 1336–1338, 1998.

[9] M. K. Goldstein, R. W. Coleman, S. W. Tu et al., "Translating research into practice: organizational issues in implementing automated decision support for hypertension in three medical centers," *Journal of the American Medical Informatics Association*, vol. 11, no. 5, pp. 368–376, 2004.

[10] D. Kalra, T. Beale, and S. Heard, *The openEHR Foundation*, vol. 115 of *Studies in Health Technology and Informatics*, 2005, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16160223.

[11] J.-F. Ethier, O. Dameron, V. Curcin et al., "A unified structural/terminological interoperability framework based on lexEVS: application to TRANSFoRm," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 986–994, 2013.

[12] J.-F. Ethier, V. Curcin, A. Barton et al., "Clinical data integration model. Core interoperability ontology for research using primary care data," *Methods of Information in Medicine*, vol. 54, no. 1, pp. 16–23, 2014.

[13] W. Kuchinke, T. Karakoyun, C. Ohmann et al., "Extension of the primary care research object model (PCROM) as clinical research information model (CRIM) for the 'learning healthcare system,'" *BMC Medical Informatics and Decision Making*, vol. 14, no. 1, article 118, 2014.

[14] S. Garde, E. Hovenga, J. Buck, and P. Knaup, "Expressing clinical data sets with openEHR archetypes: a solid basis for ubiquitous computing," *International Journal of Medical Informatics*, vol. 76, pp. S334–S341, 2007.

[15] W. Kuchinke, C. Ohmann, R. A. Verheij et al., "A standardised graphic method for describing data privacy frameworks in primary care research using a flexible zone model," *International Journal of Medical Informatics*, vol. 83, no. 12, pp. 941–957, 2014.

[16] V. Curcin, S. Miles, R. Danger, Y. Chen, R. Bache, and A. Taweel, "Implementing interoperable provenance in biomedical research," *Future Generation Computer Systems*, vol. 34, pp. 1–16, 2014.

[17] IHE, "IHE Quality, Research, and Public Health Technical Framework Supplement. Structured Data Capture," http://www.ihe.net/uploadedFiles/Documents/QRPH/IHE_QRPH_Suppl_SDC.pdf.

[18] O. Kostopoulou, A. Rosen, T. Round, E. Wright, A. Douiri, and B. Delaney, "Early diagnostic suggestions improve accuracy of GPs: a randomised controlled trial using computer-simulated patients," *British Journal of General Practice*, vol. 65, no. 630, pp. e49–e54, 2014.

Method

Our requirements were to develop a data mediation system based on a local-as-view approach which treats structural and terminological models as tightly bound and interdependent entities in order to fully present and share semantics to meaningfully use data in the LHS.

In this article published by the Journal of the American Medical Informatics Association in 2013, we present our approach to address these requirements. (Ethier *et al.* 2013) As a first step, methods to unify structural and terminological models were explored. To bridge the terminology divide, terminology servers are used to manage, curate and serve terminologies. The main open source (a TRANSFoRm requirements) systems are LexEVS and Bioportal. (Noy *et al.* 2009; LexEVS) A comparison between the two has been carried out by our team based on the requirements of a LHS. Bioportal is an extremely valuable resource, provides multiple access end-points (like SPARQL) and lowers entry costs as the main infrastructure is maintained by the National Center for Biomedical Ontologies at Stanford. (BioPortal SPARQL Query Browser) On the other hand, the centralized aspect means that a specific project cannot control the development of the platform. And while "private" ontologies can be loaded into the system, private partners in TRANSFoRm did not trust the platform enough to transfer corporate information to Bioportal. On the other hand, LexEVS is developed by a team at the Mayo clinic and used by multiple institutions. (Pathak *et al.* 2009) It can be installed and run locally giving the group complete control. It is also quite flexible. It supports multiple terminology formats, mappings between terminologies and value sets creation.

Upon further analysis, useful parallels can be drawn between terminological requirements and structural requirements in mediation systems. Both require selection of elements based on various attributes. They also similarly need to support mapping creation to align concepts. (Shvaiko and Euzenat 2005) LexEVS was therefore evaluated to verify its capacity to support both terminological but also structural resources required for the LHS. The article presents the successful method created to achieve this. Using LexEVS permits operations on both terminological and structural (and information) models using the HL7 Common Terminology Services 2 standard. (CTS2 - HL7Wiki) It also supports the binding of both models on the same platform. As an added bonus, LexEVS capabilities like multi-lingual support and versioning, essential to support the evolution of various terminologies, can also be leveraged for structural models and mappings. (Overhage *et al.* 2012)

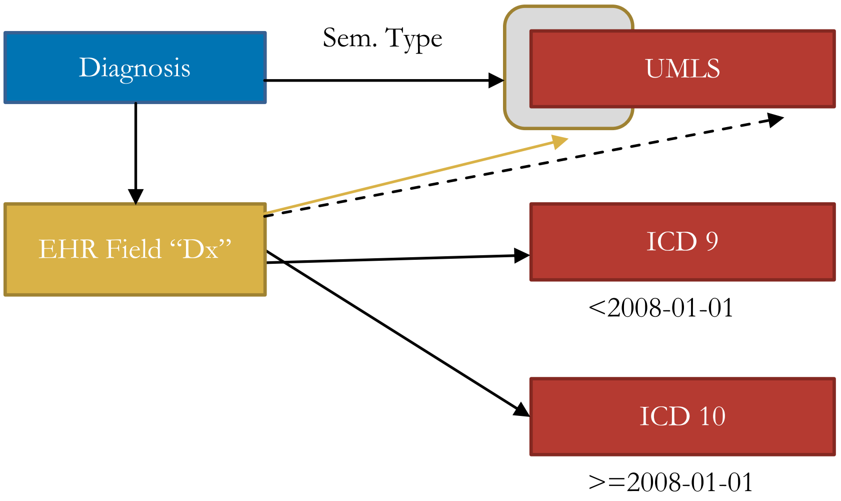


Figure 2: Structural-terminological models binding example

Figure 2 presents an example of structural-terminological bindings (structural models on the left, terminological models on the right). A concept from the mediation conceptual model (Diagnosis) is aligned with a local source database field identified as "Dx". The framework expects UMLS codes as values for diagnosis, but constrained to a subset of codes with the relevant semantic type ("Disease or Syndrome"). This constraint can then be shared with the aligned concept ("Dx" field) in the local source, meaning that an ICD 10 code of R05 (Cough) should not be present in the field since its semantic type is "Sign or Symptom". This can serve as the basis for data consistency and quality verifications. The example also illustrates that the "Dx" field contains ICD 9 codes before 2008 but ICD 10 codes afterwards.

Using the resources exposed by LexEVS, the basic unit to express the semantics of a quantum of data is a triplet: General model unique identifier | operator | terminology code(s), strings or numbers. To express a diagnosis of GORD for a patient, the following triplet could be created: OGMS_0000073;[IN];([UMLS:2015AA:C0017168]).

The first part represents the general model unique identifier referring to diagnosis, the second is the operator and the last one represents a terminological code including the terminology name, version and code (UMLS release 2015AA, code C0017168 which represents Gastro-oesophageal reflux disease). The terminology list can contain codes from various terminologies, if necessary. (Taboada *et al.* 2009) This construct also ensure tight binding of structural and terminological information at the level of the query building blocks.

A unified structural/terminological interoperability framework based on LexEVS: application to TRANSFoRm

Jean-François Ethier,[1] Olivier Dameron,[1] Vasa Curcin,[2] Mark M McGilchrist,[3] Robert A Verheij,[4] Theodoros N Arvanitis,[5] Adel Taweel,[6] Brendan C Delaney,[7] Anita Burgun[1]

[1]INSERM UMR936, Université de Rennes 1, Rennes, France
[2]Department of Computing, Imperial College London, London, UK
[3]Health Informatics Centre, University of Dundee, Dundee, UK
[4]NIVEL Primary Care Database, NIVEL, Utrecht, The Netherlands
[5]School of Electronic, Electrical and Computer Engineering, University of Birmingham, Birmingham, UK
[6]Department of Computer Science, King's College London, London, UK
[7]Department of Primary Care and Public Health Sciences, King's College London, London, UK

Correspondence to
Dr Jean-François Ethier, INSERM UMR936, Faculté de Médecine, Université de Rennes 1, 2, av. Léon Bernard, Rennes 35043, France; ethierj@gmail.com

Received 31 August 2012
Revised 10 March 2013
Accepted 14 March 2013
Published Online First 9 April 2013

To cite: Ethier J-F, Dameron O, Curcin V, *et al*. *J Am Med Inform Assoc* 2013;**20**:986–994.

ABSTRACT

Objective Biomedical research increasingly relies on the integration of information from multiple heterogeneous data sources. Despite the fact that structural and terminological aspects of interoperability are interdependent and rely on a common set of requirements, current efforts typically address them in isolation. We propose a unified ontology-based knowledge framework to facilitate interoperability between heterogeneous sources, and investigate if using the LexEVS terminology server is a viable implementation method.

Materials and methods We developed a framework based on an ontology, the general information model (GIM), to unify structural models and terminologies, together with relevant mapping sets. This allowed a uniform access to these resources within LexEVS to facilitate interoperability by various components and data sources from implementing architectures.

Results Our unified framework has been tested in the context of the EU Framework Program 7 TRANSFoRm project, where it was used to achieve data integration in a retrospective diabetes cohort study. The GIM was successfully instantiated in TRANSFoRm as the clinical data integration model, and necessary mappings were created to support effective information retrieval for software tools in the project.

Conclusions We present a novel, unifying approach to address interoperability challenges in heterogeneous data sources, by representing structural and semantic models in one framework. Systems using this architecture can rely solely on the GIM that abstracts over both the structure and coding. Information models, terminologies and mappings are all stored in LexEVS and can be accessed in a uniform manner (implementing the HL7 CTS2 service functional model). The system is flexible and should reduce the effort needed from data sources personnel for implementing and managing the integration.

INTRODUCTION

Biomedical research increasingly relies on the integration of information from multiple data sources, obtained either primarily for the purposes of research, such as trial data and genetic samples, or through secondary use of routinely collected data, for example, electronic health records (EHR). However, the heterogeneity of these data sources represents a major challenge to the research task.[1–3] Two levels of heterogeneity can be distinguished: structural and terminological. First, information models are used to represent the organization of data structures in information systems.[4–6] Variation in their forms and approaches generates structural heterogeneity of the data models. Second, numerous medical coding systems (terminologies) are used to represent diagnoses, procedures, and treatments in health databases,[7] frequently with many-to-many mappings between them, creating semantic heterogeneity, sometimes also referred to as terminological heterogeneity.[8]

Rector[8] mentions that these two types of heterogeneity, structural and semantic, are not independent as there are mutual constraints between the information models and coding systems.[9] This interdependence corresponds to what Rector calls the 'binding' between an information model and a coding system, and presents a notorious source of ambiguity in clinical systems.[4] At the time of coding, implicit knowledge is sometimes used but not formally represented in the information model. Some models function under the closed world assumption, whereby omission implies falsehood, while others support the open world assumption in which omission merely states that the information is not available. Further complexity is caused by differences in granularity, depth, coverage and composition (single term vs expressions) between models.

This article proposes a unified framework for the integration of heterogeneous information models and terminologies to construct a single solution for structural and semantic interoperability. This approach is currently being adopted in TRANSFoRm, a EU FP7 project that aims to support the integration of clinical and translational research data comprehensively in the primary care domain.[10 11]

BACKGROUND AND SIGNIFICANCE

Structural and semantic interoperability in biomedical data has been explored in a number of initiatives. Given our interest in translational medicine and data reusability, we focus here on those allowing federated queries from multiple clinical repositories and EHR.

There have been attempts to create generic information models to serve as standards, including the OpenEHR reference model, the informatics for integrating biology and the bedside (i2b2) model, the HL7 reference information model and the clinical data acquisition standards harmonization (CDASH).[12–15] An ongoing international collaboration between standards organizations and industry partners, the clinical information modeling initiative, aims at bringing together a variety of approaches to clinical data modeling (HL7 templates, openEHR archetypes, etc) as a series of underlying reference models.[16] A similar endeavor is ongoing with the biomedical research integrated domain group in the research area.[17] Nevertheless, many existing data sources are not designed according to these initiatives.

Approaches to structural heterogeneity can be grouped in two categories: extract-transform-load (ETL) systems and mediators systems. In the former, the different data sources to be integrated (eg, data warehouses) are all expected to conform to some structural model. This is achieved by carrying out an ETL process on an existing relational database to transfer the data into a single target model. Multiple projects have been built on this approach. The shared health research information network (SHRINE) aims at bringing together various i2b2 clinical data repositories.[13 18 19] The i2b2 model is also used by other projects like TRANSMART.[20] The Stanford translational research integrated database environment, an initiative from Stanford, uses the HL7 reference information model as a foundation for their model while EU-ADR developed its own common model.[21 22] Finally, the electronic primary care research network (ePCRN) project, focusing on the primary care domain, based its structure on the American Society for Testing and Materials continuity of care record information model.[23 24]

Other systems use a mediator approach to address structural heterogeneity. Some central schema is mapped to the local schemas of individual data sources, which retain their original structure. These central schemas were initially described as ontologies.[25] Projects such as advancing clinico-genomic trials in the cancer domain leveraged this approach.[26] Other projects implemented mediators in different ways. The biomedical informatics research network (BIRN) and its follow-up initiative the neuroscience information framework are using an XML approach.[27–29] The cancer biomedical informatics grid (caBIG) is a long-standing National Cancer Institute (NCI)-driven initiative to federate healthcare data with sources represented as unified modeling language (UML) models.[30–32] A similar modeling approach is used by the federated Utah research and translational health e-repository (FURTHeR) and electronic health record for clinical research.[33 34] None of these implementations use vocabulary services to support their structural aspects.

The terminological needs of various projects are handled internally. The SHRINE project uses a pivot terminology and BIRN stores term mappings in a relational database.[35 36] The smart open services for European patients (epSOS) project is developing an ontology to address the multilingual and mapping needs of its community.[37 38] Nevertheless, terminology servers are often involved like Apelon DTS in FURTHeR and Bioportal in ONCO-I2B2.[39 40]

The LexEVS terminology server, having originally been developed in the context of the caBIG initiative, is being used by several projects (eg, ePCRN, NCI thesaurus browser).[24 41 42] The web-based server bioportal also uses it as part of its infrastructure.[43] LexEVS permits unification of all loaded terminologies under the LexGrid format (including ontologies expressed as ontology web language).[44] It allows a range of deployment options, from a local installation to a grid service, and is available under an open source license. V.6 of LexEVS implements the HL7 common terminology services 2 (CTS 2) service functional model (SFM), although it does not conform to the HL7 CTS 2 OMG specification because the specification was finalized after V.6 was released.[45 46] Prior to our efforts, LexEVS implementations have mostly been used to support terminological information.

Binding between information models and terminologies presents a challenge in its own right. A number of projects mentioned above have developed their own solutions; nevertheless, standards for metadata registries have been created to address this question (eg, ISO 11179).[47] Projects such as eMERGE and caBIG use the cancer data standard repository (caDSR).[48] It stores data elements described by a definition of what is represented as well as the list of valid values. caBIG binds its UML models with the terminologies through use of these data elements. eMERGE also uses the caDSR to harmonize local genotype and phenotype data elements. The binding of structure and terminology has also been addressed in the context of HL7 with the TermInfo initiative currently focusing on the use of SNOMED CT in HL7 V3.[49]

All of these projects consider structural and semantic aspects of interoperability to be distinct, leading them to be managed separately, although the separation between structure and terminology is drawn differently in different projects. Recognizing their dependencies and that terminological and structural operations share a common set of requirements (through binding and mappings), we hypothesized that a unified ontology-based knowledge framework can facilitate interoperability between heterogeneous sources, without having to create a separation and different tools for management. Based on our analysis of terminological solutions, we investigated whether LexEVS was a functional tool to implement this approach.

In the next section, we present the framework and describe the generic approach for each of its components. We then test this method on a clinical study example from the TRANSFoRm project, focusing on integrating two primary care data repositories, the NIVEL primary care database (NPCD)[50] of the Netherlands Institute for Health Services Research (NIVEL)[51] and the general practice research database (GPRD)[52] of the UK's Medicines and Healthcare Products Regulatory Agency.[53]

MATERIALS AND METHODS

The main aims of our work are to simplify the handling of heterogeneous data sources for the users and to minimize the interoperability implementation workload for the data sources. We believe the mediation paradigm best meets these goals.[25] Instead of using ETL to enforce a uniform information model, our framework uses mappings to relate local models to a general information model (GIM). This also facilitates user operations as they only need to interact with the general model and do not need to be familiar with each data source's information model.

The mediation framework has been constructed according to the local-as-view principle.[54] In this approach, each source schema is defined as a set of views on the global schema, as opposed to the global-as-view principle in which the global schema is defined in terms of the sources. So the GIM does not have to be derived directly from any source. Rather, it should be built to construct a sound and logical view of the domain of interest in order to make sure all required concepts are present. This ensures scalability, as adding a new source does not necessitate a modification of the GIM. It also presents a more stable model to the user.

In our framework, GIM is represented as an ontology, allowing it to be stored in the LexEVS terminology server together with the data source models (DSM) and the terminologies. Mappings between GIM and data sources can then be uniformly created, stored and leveraged as described below. In parallel, similar methods can be used to handle terminological operations.

Architecture overview

The modeling infrastructure resides entirely within a terminology server, enabling unification of structural and semantic modeling and operations within this server. Several types of models are present:

1. The GIM
2. Models describing each data source (DSM)
3. Mapping sets between the sources and the GIM—one set per source
4. Terminologies used to code the data elements (eg, International Classification of Diseases (ICD)-10 codes…)
5. Mappings between terminologies.

An overview of how the different models interact together is presented in figure 1, which shows a user query being sent to mulitple data sources. Security and other administrative issues have been intentionally left out of this list in order to focus on the relevant steps for this demonstration.

1. The query is expressed using GIM concepts
2. The mediation engine generates a specific query for each data source
3. The data sources fulfill the requests
4. The returned dataset has its structure aligned with the GIM
   - DSM to extract which terminology was used to code a given concept in the source
5. If possible and desired, the system can semantically align resulting coded values based on the terminologies used by one of the sources or a separate terminology. This operation uses:
   - Terminologies and mappings between terminologies to transcode the values.

General information model

The GIM is used to represent a unified view of the domain concepts and their relationships. For example, date of birth, diagnosis and patient are all relevant concepts in a clinical care context. Each concept also has intrinsic properties. Given the data integration function of the ontology and its role as a mediation schema, we chose a realist approach using basic formal ontology (BFO) 1.1 as the foundation of the model.[55] [56] The implementation of BFO as a formal, description logics ontology allows easier interaction with projects using semantic web technologies (like epSOS), or other parts of projects implementing the framework. For example, the provenance service and the decision support service from TRANSFoRm both rely on ontologies and will need to interact closely with the unified integration framework.

Figure 2 illustrates how 'gender' and its relevant attributes represented in GIM are rendered once loaded in LexEVS.

The 'codedWith' properties of the concept support binding between the information model and the relevant terminology (or value set) and contribute to its semantics representation. In this case, it indicates that values for this concept are to be represented with the terminology named 'gim_gender' stored in LexEVS. Multilingual capabilities are handled natively within LexEVS by combining property values with a language descriptor. When a translation is provided, this allows the model also to propose a multilingual solution without resorting to another system.

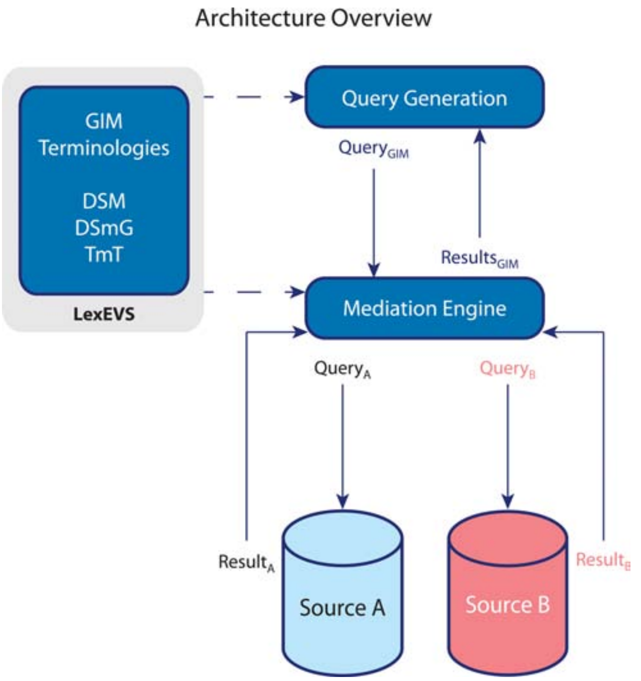


Figure 1 Architecture supporting model interactions based on LexEVS for query mediation-based query resolution.

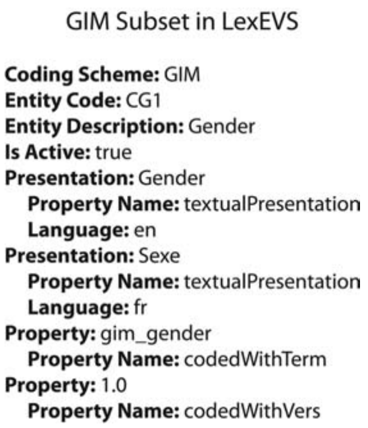
Data source models

A new DSM is defined for every data source to be supported. The goal of this stage is to provide enough information to the system in order to translate a query based on the GIM into the local language used to query the source. The exact nature of the properties and relations will be related to the underlying type of source to be modeled.

For example, a SQL data source 'SA' would have hierarchical relations such as hasTable and hasField with other relations representing the relations between the tables (oneToMany, OneToOne…) with the keys on each side. Another data source 'SB' could be an XML document, with XPath as its query language. A model fulfilling the same goal can be created describing nodes, elements and attributes.

A DSM fragment is illustrated in figure 3, representing a field. In terms of concept properties, we have some similarities with the GIM but also specific properties for a SQL source concept.

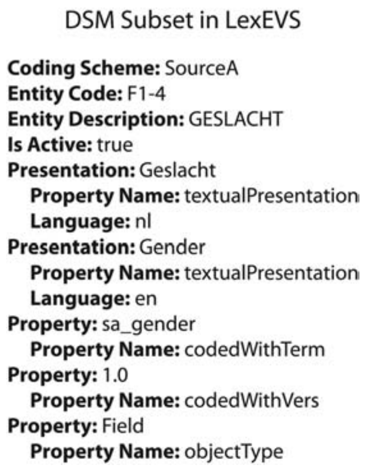
The objectType property gives the nature of the concept (field) while the name of the object is in the description. Multiple textual presentations (here Dutch and English) can be



GIM Subset in LexEVS

Coding Scheme: GIM
Entity Code: CG1
Entity Description: Gender
Is Active: true
Presentation: Gender
  Property Name: textualPresentation
  Language: en
Presentation: Sexe
  Property Name: textualPresentation
  Language: fr
Property: gim_gender
  Property Name: codedWithTerm
Property: 1.0
  Property Name: codedWithVers

Figure 2 General information model—partial representation of 'gender' attributes in LexEVS.



DSM Subset in LexEVS

Coding Scheme: SourceA
Entity Code: F1-4
Entity Description: GESLACHT
Is Active: true
Presentation: Geslacht
  Property Name: textualPresentation
  Language: nl
Presentation: Gender
  Property Name: textualPresentation
  Language: en
Property: sa_gender
  Property Name: codedWithTerm
Property: 1.0
  Property Name: codedWithVers
Property: Field
  Property Name: objectType

Figure 3 Data sources models—partial representation in LexEVS of a field named 'GESLACHT' from a source SQL database.

created to provide translations in order to facilitate the use of the information in multiple contexts. As with the GIM, 'codedWith' properties hold the name and versions of the terminology (or local value set) used to code data for this concept (a field in this example). Note that this does not need to be the same terminology in all DSM and GIM. This allows a DSM to register the specific terminology (or value set) used to code the information locally, irrespective of what is registered with GIM.

Mappings between a source and the GIM

A mapping set does not need to duplicate the concepts from the model but simply reference them via their code and coding scheme name. A relation is then created for each correspondence between a GIM concept and a DSM concept.

We developed a generic mapping model defining data transformation operations to align source data values with the GIM, supporting not only one-to-one mappings but also more complex cases. One-to-one operations include simple mappings such as a date corresponding to a date/time value, while a more complex case would consist of two distinct but related fields. For example, a symptom (a code from a terminology) can possibly denote multiple entity types (in GIM). For example, 'abdominal pain' can be used to code a 'presenting complaint', a 'symptom' or even sometimes a 'final diagnosis' if no clear diagnosis emerges during the consultation. Some data sources, instead of having three fields representing the three possible entity types, will have two fields: one storing the actual symptom code and one for the entity type. For example, field A would store the value 'abdominal pain', while field B would store the entity type 'presenting complaint' in the same record, to distinguish it from someone with a diagnosis of abdominal pain as part of their medical history.

In this case, instead of linking directly from the source to the GIM, an intermediate concept is created in the mapping set. This intermediate concept will hold the condition for this relation to be true. So, if our example maps to some concept AP154 in GIM, the mapping would proceed as Field A→Condition 1 (Field B='Value 1')→GIM AP154, that is, Field A represents GIM concept AP154 only if Field B='Value 1'. Intermediate concepts can also be chained in order to combine different operations.

The model also supports the creation of a virtual element to capture implicit knowledge. For example, it could represent a laboratory unit that might not be physically present in the data source because it is always the same in the context of that source. Similarly, the mapping model can support yes/no fields (eg, a column denoting the presence or absence of diabetes), which combines both the structural and terminological elements.

Terminologies

The UMLS presents a unified view of a large number of relevant biomedical terminologies.[57] It includes over two million concepts from various vocabularies and millions of relationships. By using concept unique identifiers—used to relate codes in different terminologies but with a similar meaning—and semantic groups, it facilitates terminology alignment. The UMLS can be loaded directly in LexEVS 6, which supports all its features.

Additional LexEVS loaders are easily created to load terminologies that are not yet supported. This was exemplified by the creation of a loader for the anatomical therapeutic chemical classification system (ATC 2011) in collaboration with the LexEVS developers.

Mappings between terminologies

Once terminologies are loaded in LexEVS, mappings between them can be created in a similar way as for the data models. For some of them, relationships are readily available and can be simply loaded into LexEVS. This is typically the case for terminologies integrated in the UMLS.

For others, local mappings have to be created. For example, if a hospital uses a local coding set to identify its laboratory tests, it could be loaded into LexEVS. Subsequently, mappings between this local set and logical observation identifiers names and codes could be created. This would allow translations from the local site to a more standard terminology, thereby facilitating interoperability with other groups without having to recode data locally or create a duplicate data warehouse.

When more than two terminologies are used, mapping sets can be created between each of them or only to some selected central (pivot) terminology, which then acts as a hub for translating concepts. A pivot terminology is optional in the GIM framework and left for the users to decide on. In the absence of a designated terminology, the user can choose one of the terminologies supported in the selected sources to which the others will attempt to map.

RESULTS

The first implementation of GIM was realized as part of the EU FP7 TRANSFoRm project, which aims at supporting patient safety through integration of clinical and research settings, workflows and data.[11] The technology developed can facilitate the interactions with individual EHR systems for trial recruitment and follow-up, as well as diagnostic support. The TRANSFoRm project also relies on a workbench to explore clinical and research data repositories. To achieve this, significant challenges need to be overcome in the areas of interoperability and methods for data integration.

Clinical data integration model: GIM instantiation in TRANSFoRm

The clinical data integration model (CDIM) is the GIM instantiation in TRANSFoRm, and covers concepts relevant to data integration in primary care research such as medication, diagnosis, and laboratory tests. It is implemented as an ontology web language ontology based on the BFO 1.1.[56] It imports the general medical science,[58] the vital sign ontology[59] and the information artifact ontology.[60] The ontology also integrates concepts from existing ontologies such as the ontology for

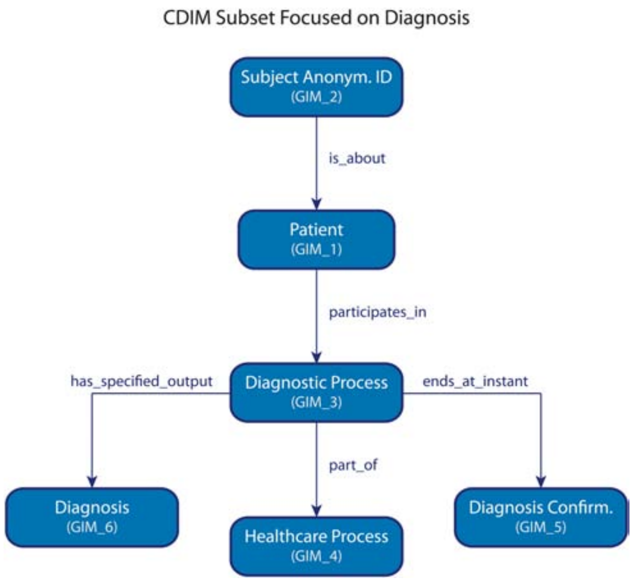


Figure 4 Clinical data integration model subset focused on diagnosis. Identifiers are in parentheses.

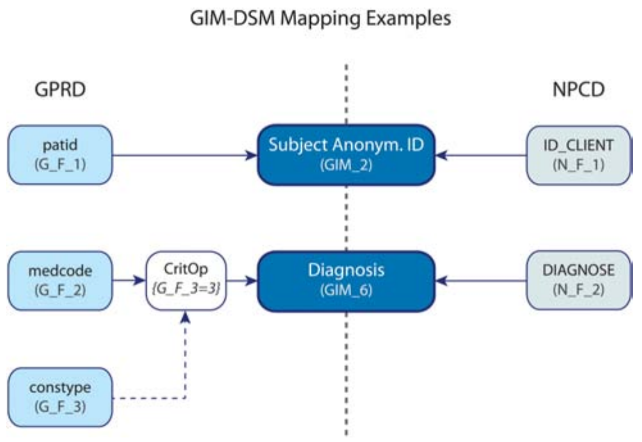


Figure 5 Mapping examples between general information model and data sources models (general practice research database and NIVEL primary care database). Identifiers are in parentheses.

biomedical investigations,[61] the gene ontology[62] and the translational medicine ontology[63] when possible.

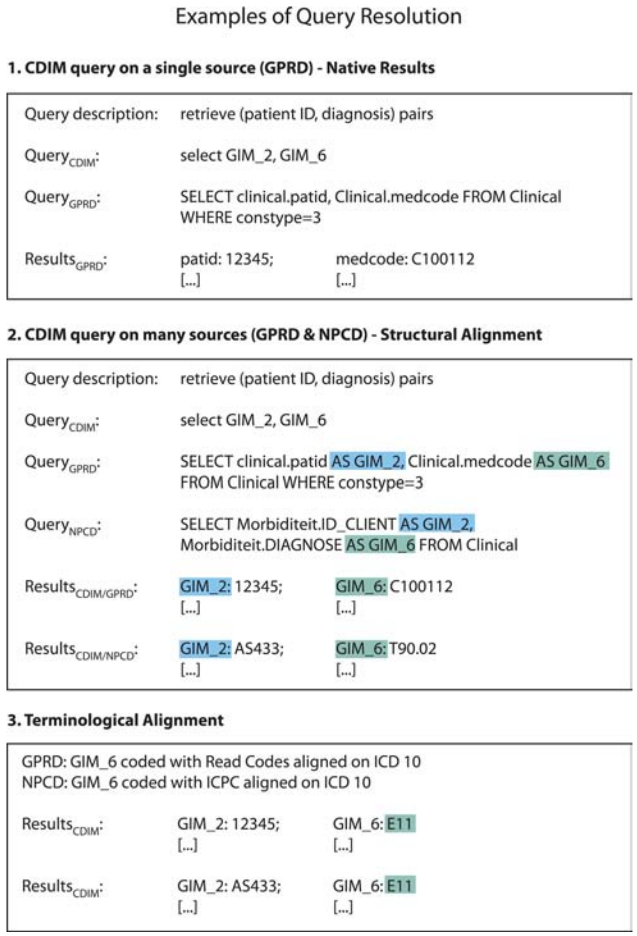
The resulting ontology has 457 classes (102 unique to CDIM) and 73 object properties (1 sub-property unique to CDIM). Twenty-one novel CDIM classes had to be introduced to represent and manage temporal aspects necessary in TRANSFoRm. All required concepts, as defined by use cases, could be modeled in CDIM. Figure 4 presents a subset of CDIM adapted to illustrate a subset of queries related to the diagnosis of diabetes.

Instantiation of structural models, terminologies and mappings

Two clinical data repositories were used to evaluate the suitability of the framework for the project: NPCD from the Netherlands and GPRD from the UK. Both their structures and the terminologies used to code information are different. For example, medication is coded with the British national formulary (BNF) codes in GPRD but the ATC classification is used in NPCD, with diagnoses coded with read codes V.2 in GPRD and ICPC V.1 in NPCD.

Structural models in XML were created for both sources using a semi-automated tool and then loaded into LexEVS. The NPCD database extract we used contained 60 521 anonymized patient records, whereas the GPRD extract made available for the project contained 5000 patient entries. Eight tables (181 fields) in NPCD and 10 tables (107 fields) in GPRD were considered in the structural models.

CDIM was mapped with 44 elements in NPCD and 47 in GPRD. High level classes such as 'processual entity' are part of CDIM and are essential to knowledge modeling but are not expected to be used as mapping targets as they are too generic. Twenty-nine mappings (32%) were one-to-one direct relations between CDIM concepts and a data source structural element. The other mappings included concatenation operations and conditional mappings (including related tables). No virtual elements were necessary for the current data source mappings. Figure 5 illustrates an example of a conditional mapping. Precise and comprehensive knowledge of each data source and its real-life usage was essential to achieve satisfactory mappings and query results. Not all fields of the data sources are targets for mappings, nor are all concepts in CDIM mapped to each data source; their coverage typically differs from CDIM. Nevertheless, all the relevant entities for the use cases were successfully mapped. Figure 5 presents those mappings necessary to illustrate the examples in figure 6.


Examples of Query Resolution

1. CDIM query on a single source (GPRD) - Native Results

Query description: retrieve (patient ID, diagnosis) pairs

$\text{Query}_{\text{CDIM}}$: select GIM_2, GIM_6

$\text{Query}_{\text{GPRD}}$: SELECT clinical.patid, Clinical.medcode FROM Clinical WHERE constype=3

$\text{Results}_{\text{GPRD}}$: patid: 12345; [...] medcode: C100112 [...]

2. CDIM query on many sources (GPRD & NPCD) - Structural Alignment

Query description: retrieve (patient ID, diagnosis) pairs

$\text{Query}_{\text{CDIM}}$: select GIM_2, GIM_6

$\text{Query}_{\text{GPRD}}$: SELECT clinical.patid AS GIM_2, Clinical.medcode AS GIM_6 FROM Clinical WHERE constype=3

$\text{Query}_{\text{NPCD}}$: SELECT Morbiditeit.ID_CLIENT AS GIM_2, Morbiditeit.DIAGNOSE AS GIM_6 FROM Clinical

$\text{Results}_{\text{CDIM/GPRD}}$: GIM_2: 12345; [...] GIM_6: C100112 [...]

$\text{Results}_{\text{CDIM/NPCD}}$: GIM_2: AS433; [...] GIM_6: T90.02 [...]

3. Terminological Alignment

GPRD: GIM_6 coded with Read Codes aligned on ICD 10
NPCD: GIM_6 coded with ICPC aligned on ICD 10

$\text{Results}_{\text{CDIM}}$: GIM_2: 12345; [...] GIM_6: E11 [...]

$\text{Results}_{\text{CDIM}}$: GIM_2: AS433; [...] GIM_6: E11 [...]

Figure 6 Examples of query resolution as applied to TRANSFoRm using clinical data integration model (figure 4), its mappings to the data sources models (figure 5) and terminologies. Highlighted segments represent each level-specific addition based on information from the models served by LexEVS.

Based on our use case and available data sources, we focused on ICD-9 and ICD-10 codes, International Classification of Primary Care (ICPC) V.1 codes, read codes V.2 for diagnoses, the ATC, the BNF for drugs, as well as on logical observation identifiers names and codes for laboratory tests.

Evaluation

We evaluated the applicability of the GIM approach to TRANSFoRm's clinical trial use cases. We focused on the retrospective diabetes cohort study.[64] This use case aims at identifying eventual associations between single nucleotide polymorphism and diabetes complications or responses to oral antidiabetic drugs. Twenty-six relevant queries were identified and were all successfully implemented, in conjunction with appropriate terminological values. For example:

- Patients ≥35 years old
  AND
  ((with a diagnosis of diabetes accompanying a prescription or an episode of care)
  OR (taking metformin OR a sulfonylurea medication in last 5 years)
  OR (having a laboratory test of glycosylated hemoglobin >6.5%
    OR a random glucose >11.0 mmol/l
    OR a fasting glucose >7.0 mmol/l))

Figure 6 demonstrates different features of the LexEVS implementation of the framework. The first example illustrates how to create the local source query based on information contained within CDIM and the DSM. The latter would contain field and table relations required to derive the SQL statement. By utilizing the mappings shown in figure 5, the query is translated in the local source query format.

Similar principles can be applied for multiple sources, but as shown in the first example of figure 6, the resulting dataset structure is based on the local source. In the example, it is not clear that 'DIAGNOSE' and 'medcode' carry a similar meaning, especially as this equivalence is only true if a condition on the field 'constype' is applied. By adjusting the local query to maintain a reference to CDIM, the resulting datasets from two data sources (NIVEL and GPRD) can be assembled in a coherent structure as in example 2.

Although both result sets now share an identical structure, the terminologies used to code the information are different. In some situations, alignment might not even be possible, at least not in a completely automated fashion as with ATC and BNF for medication types. In this diabetes example, we consider the 'coded with' properties in the local DSM, as previously described. For GPRD, 'Non-insulin dependent diabetes mellitus' in read codes V.2 can be related to an ICD-10 code (E11) by following mappings in LexEVS. The same can be done for NIVEL with ICPC-1 code T90.02 to ICD-10 code E11. The final unified dataset is homogenous and consistent semantically as in example 3.

DISCUSSION

Achieving interoperability between health data sources such as EHR and registries is a challenging but crucial endeavor for both designers and users of health information technology. The structural and terminological aspects of data source interoperability, while intrinsically linked, have traditionally been handled separately.[65 66] From a structural perspective, a number of projects have adopted a common model to which each source is expected to comply, whether when inputting data (eg, CDASH in the clinical research domain)[15] or when data are being extracted (eg, EU-ADR focusing on adverse event analysis).[22]

Other projects have opted for a mediation approach, with a centralized knowledge model, often represented as an ontology. XML and UML designs are also possibilities, as utilized in the BIRN and FURTHeR projects, respectively. Our framework is built around GIM as the central knowledge model, expressed as an ontology with a realist approach based on BFO 1.1.

The semantic challenges are addressed either through dedicated project-specific tools or through terminological servers, such as the one used in the ePCRN project. The GIM framework is novel in that it uses a terminological server not only for handling semantic interoperability, but for structural aspects as well.

Binding both terminological and structural aspects, when they are managed separately, is a challenge that has previously been handled through the use of metadata registries such as caDSR, as used in the caBIG and eMERGE projects.[30 67] The registries allow data elements to be created in which a definition and a list of permissible values is attached. Our framework avoids this situation by handling the binding in the mediation structure, in which both sets of models are located already. It allows data elements present in existing data sources to be described and integrated readily in the context of GIM and allows the use of local code value sets easily as they are stored in the framework.

Our approach represents a step beyond the traditional interoperability paradigm involving a different set of tools for dealing with structural, terminological and binding challenges, in that we present a unified framework that provides an integration solution for these facets inside a single structure. Our LexEVS implementation of GIM, as demonstrated in the TRANSFoRm project, allows a query to be expressed using clinical concepts from a single generic model that is represented as an ontology, and allows its translation into source-specific queries, which then return the results from each source, simplifying and standardizing the interoperability task.

Strengths and limitations

One of the biggest barriers to the usage of federated data sources is the resource and effort expected from the data sources to participate in a collaborative structure.[3] In order to mend heterogeneity between two data sources, related elements must be mapped to each other. Whether structural models, such as database schemas, or terminologies are to be aligned, the processes share a common subset of requirements.[68] Multiple approaches have been developed to address the issue.[57 69] Our infrastructure does not necessitate a priori substantial changes to the structure of the data source. If desired, ETL may be used to transform the initial data schema into a derived schema closer to GIM, and this could facilitate the use of direct mappings. If an organization already has a data warehouse, it might be used as is, thereby reducing integration effort and avoiding data duplication.

The architecture presented decouples the interoperability modeling aspects from the application itself. For some data sources, especially EHR, exposing the structure of their databases might not be possible or desirable. In this case, an instance of LexEVS can be installed on a local server, allowing query translation to happen at the local level.

From the maintenance perspective, the addition of a new piece of information to a source will necessitate mappings to the relevant GIM terms before becoming usable.[9] Note that our approach can leverage the GIM semantic richness to make this mapping step easier.[70] This occurred with the CDIM implementation of GIM in the TRANSFoRm project, in which we use 'codedWith' properties to suggest concepts that might share similar semantics. Similarly, distance between concepts in the graph can be used to suggest related concepts. Mappings within TRANSFoRm are currently created manually but should it be expanded, mapping tools will be required in order to support its development. Our LexEVS implementation supports most attributes necessary to allow such work.[70] This has recently been identified as a core challenge to the field by Shvaiko and Euzenat,[68] and we believe that our approach can contribute to an alignment infrastructure, fostering collaboration.

There are a number of advantages to using LexEVS as the implementation technology. The GIM ontology is stored in the LexEVS terminology server, allowing us to leverage its two optimization axioms: 'fully restrict then query' and 'lazy loads'. The former minimizes resource requirements by allowing the system to restrict any query fully, including operations on sets (eg, intersections, unions or differences) before running it against the data source. The latter technique preferentially loads only certain types of information in the first pass while retaining a pointer to load more information dynamically should this be needed. Together, these facilitate efficient query mediation on heterogeneous data sources.

Our approach also benefits directly from LexEVS capabilities for handling versioning. Multiple versions of the models, terminologies and mappings can coexist in the system, and be maintained independently from our framework, removing the need for a separate implementation of versioning. Similarly, multilingual capabilities supported by LexEVS can be used for many operations without resorting to an ancillary tool.

Once loaded and functional, the framework can leverage intrinsic capabilities of LexEVS to create value sets (ie, subsets of related concepts), which can then be used to handle terminological needs (eg, codes used to represent drugs to treat diabetes) and manage GIM concept groups. For example, relevant concepts related to laboratory tests can be grouped in order to facilitate searching and browsing. This is different from other efforts in which structural models are stored in project-specific structures. Using LexEVS to manage GIM and DSM automatically provides the methods that implement the HL7 CTS 2 SFM, and ultimately HL7 CTS 2 OMG, ensuring that the implementation remains maintainable and reusable.[71]

The level of automation for query translation and results aggregation depends on the possibility of creating meaningful mappings between relevant terms.[72 73] We showed in our example that mappings between different terminologies can be utilized to automate the process fully for some situations. Nevertheless, some terminology pairs do not lend themselves to such an exercise. These include the ATC and BNF terminologies for therapeutic substances.[74 75] Their approach to classification varies in granularity, depth and coverage, leading for some terms to one-to-many mappings or absence of related concept. In such a scenario, the infrastructure can readily support a user interface in which similar, but not necessarily equivalent, terms in different terminologies used by different sources could be suggested, edited and finally approved by the user instead of being automatically chosen.

Applicability

The infrastructure is currently being deployed in the pan-European TRANSFoRm project, with a view to deploying it in other EU and US translational research projects in academia and industry. Specific TRANSFoRm activities that require combined semantic and structural integration include:

- Support for dynamic and persistent linkage between data sources for widely scalable epidemiological studies.
- Support for clinical decision support embedded in the EHR, enabling capture and recording of clinical diagnostic cues in a controlled form.
- Support for real time linkage to a variety of different EHR systems for extraction of clinical data elements into an electronic case report form and write-back of controlled data elements to the EHR to serve as an eSource for regulated clinical trials.

Deploying CDIM as a unified framework in this setting allows the project tools to have full control over the content and structure of queries sent to data sources, and demonstrated its applicability to multiple deployment scenarios, including distributed installations. This study showed that this unified framework, supported by LexEVS, is a suitable platform in which to achieve these tasks in the context of two exemplar databases. The tool chosen in TRANSFoRm was LexEVS. Nevertheless, in a different context, other tools such as Bioportal might also have the potential to support the framework.

CONCLUSION

In this paper, we presented a novel, unifying approach to address interoperability challenges in heterogeneous data sources, by representing structural and semantic models in a single framework. This represents a significant departure from the previous strategies for addressing interoperability in translational research, and it has been successfully demonstrated within the context of the clinical research studies of the EU TRANSFoRm project.

The advantage of this approach is that the systems using the architecture can rely solely on GIM concepts, abstracting over both the structure and coding specificities of the data sources. Information models, terminologies and mappings are all stored in LexEVS and can be accessed using the same methods (implementing the HL7 CTS 2 SFM). The system is flexible, and should reduce the integration effort required from the data sources, thereby lowering the cost of entry of this type of research for smaller institutions, and removing the need for larger institutions to invest in additional data warehousing.

Acknowledgements The authors would like to thank their colleagues from the TRANSFoRm project for their support and insightful discussions regarding this endeavor. In particular, Peter Leysen, Hilde Bastiaens, and Anna Nixon Andreasson developed the use cases. Nasra Khan provided data extracts and processing at NIVEL, and Nick Wilson at CPRD. Jean-Karl Soler assisted in terminology mappings. Lei Zhao developed terminology services in TRANSFoRm. The authors also thank the members of the LexEVS development team for their invaluable help.

Contributors JFE created, designed and tested the generic framework, created the ontology and drafted the manuscript. AB and OD participated in the design of the generic framework and worked on the ontology development. AB, OD, VC and BCD helped to draft the manuscript. AT, TNA and BCD contributed to the conception and design of the TRANSFoRm project and its architecture. JFE, VC, AB, MMM, TNA and AT participated in the implementation of the framework in TRANSFoRm. JFE, AB and TNA contributed to the TRANSFoRm vocabulary services. MMM, JFE and RAV worked on the TRANSFoRm data sources model structures. All authors critically reviewed the manuscript and approved the final version.

Funding This work was supported in part by the European Commission—DG INFSO (FP7 247787).

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Cimino J J. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf Med* 1998;37:394–403.

2 Cimino JJ. In defense of the Desiderata. *J Biomed Inform* 2006;39:299–306.

3 Sujansky W. Heterogeneous database integration in biomedicine. *J Biomed Inform* 2001;34:285–98.

4 Rector AL, Qamar R, Marley T. Binding ontologies and coding systems to electronic health records and messages. *Appl Ontol* 2009;4:51–69.

5 Eichelberg M, Aden T, Riesmeier J, *et al*. A survey and analysis of electronic healthcare record standards. *ACM Comput Surv* 2005;37:277–315.

6 Haux R, Kulikowski C, editors. IMIA Yearbook of Medical Informatics 2006. *Methods Inf Med* 2006;45(Suppl 1):S136–44.

7 Geissbuhler A, Kulikowski C, editors. IMIA Yearbook of Medical Informatics 2008. *Methods Inf Med* 2008;47(Suppl 1):67–79.

8 Rector AL. Clinical terminology: why is it so hard? *Methods Inf Med* 1999;38:239–52.

9 Qamar R, Kola JS, Rector AL. Unambiguous data modeling to ensure higher accuracy term binding to clinical terminologies. *AMIA Annu Symp Proc* 2007;2007:608–13.

10 Delaney B. TRANSFoRm: translational medicine and patient safety in Europe. In: Grossman C, Powers B, McGinnis JM, eds. *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: 198–202.

11 TRANSFoRm Project. http://www.transformproject.eu (accessed 11 Apr 2012).

12 Beale T, Heard S, Kalra D, *et al*. The openEHR reference model—EHR information model—Release 1.0.2. http://www.openehr.org/releases/1.0.2 (accessed 29 Jun 2012)

13 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;2007:548–52.

14 Schadow G, Mead CN, Walker DM. The HL7 reference information model under scrutiny. *Stud Health Technol Inform* 2006;124:151–6.

15 CDASH—Basic Recommended Data Collection Fields for Medical Research. http://www.cdisc.org/cdash (accessed 8 Dec 2012).

16 Clinical Information Modelling Initiative. http://www.openehr.org/326-OE.html?branch=1&language=1 (accessed 8 Dec 2012).

17 Fridsma DB, Evans J, Hastak S, *et al*. The BRIDG Project: a technical report. *J Am Med Inform Assoc* 2008;15:130–7.

18 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;16:624–30.

19 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;17:124–30.

20 Szalma S, Koka V, Khasanova T, *et al*. Effective knowledge management in translational medicine. *J Transl Med* 2010;8:68.

21 Lowe HJ, Ferris TA, Hernandez PM, *et al*. STRIDE—an integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc* 2009;2009:391–5.

22 Avillach P, Dufour J-C, Diallo G, *et al*. Design and validation of an automated method to detect known adverse drug reactions in MEDLINE: a contribution from the EU-ADR project. *J Am Med Inform Assoc* 2013;20:446–52.

23 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;10:54–9.

24 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;19:93–7.

25 Wiederhold G. Mediators in the architecture of future information systems. *Comput J* 1992;25:38–49.

26 Martin L, Anguita A, Graf N, *et al*. ACGT: advancing clinico-genomic trials on cancer—four years of experience. *Stud Health Technol Inform* 2011;169:734–8.

27 Gupta A, Ludascher B, Martone ME. Knowledge-based integration of neuroscience data sources. In: *Scientific and Statistical Database Management, 2000. Proceedings. 12th International Conference*, 2000: 39–52.

28 Astakhov V, Gupta A, Grethe JS, *et al*. Semantically based data integration environment for biomedical research. In: *Proc of the 19th IEEE Symp Comput Based Med Syst*; 22–23 June 2006, Washington, DC: IEEE Computer Society, 2006:171–6.

29 Ashish N, Ambite JL, Muslea M, *et al*. Neuroscience data integration through mediation: an (F)BIRN case study. *Front Neuroinform* 2010;4:118.

30 Stanford J, Mikula R. A model for online collaborative cancer research: report of the NCI caBIG project. *Int J Healthc Technol Manag* 2008;9:231–46.

31 González-Beltrán A, Tagger B, Finkelstein A. Federated ontology-based queries over cancer data. *BMC Bioinformatics* 2011;13(Suppl. 1):S9.

32 Saltz J, Oster S, Hastings S, *et al*. caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. *Bioinformatics* 2006;22:1910–16.

33 Livne O, Schultz N, Narus S. Federated querying architecture with clinical & translational health IT application. *J Med Syst* 2011;35:1211–24.

34 Ouagne D, Hussain S, Sadou E, *et al*. The Electronic Healthcare Record for Clinical Research (EHR4CR) information model and terminology. *Stud Health Technol Inform* 2012;180:534–8.

35 Core Ontology—SHRINE. https://open.med.harvard.edu/display/SHRINE/Core+Ontology (accessed 18 Apr 2012).

36 Bug W, Ascoli G, Grethe J, *et al*. The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics* 2008;6:175–94.

37 D3.5.2_Appendix_E_Ontology_Specifications_01.pdf. http://www.epsos.eu/uploads/tx_epsosfileshare/D3.5.2_Appendix_E_Ontology_Specifications_01.pdf (accessed 8 Dec 2012).

38 epSOS: About epSOS. http://www.epsos.eu/home/about-epsos.html (accessed 11 Apr 2012).

39 Matney S, Bradshaw R, Livne O, *et al*. Developing a semantic framework for clinical and translational research. In: *AMIA Summit on Translational Bioinformatics; 7–9 March 2011*, Bethesda, MD: AMIA, 2011:24.

40 Segagni D, Tibollo V, Dagliati A, *et al*. An ICT infrastructure to integrate clinical and molecular data in oncology research. *BMC Bioinformatics* 2012;13(Suppl. 4):S5.

41 NCI Thesaurus Browser. https://cabig-stage.nci.nih.gov/community/tools/NCI_Thesaurus (accessed 8 Dec 2012).

42 LexEVS 6.0 Architecture. https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexEVS_6.0_Architecture (accessed 30 May 2011).

43 Noy NF, Shah NH, Whetzel PL, *et al*. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009;37(web server issue):W170–3.

44 Pathak J, Solbrig HR, Buntrock JD, *et al*. LexGrid: a framework for representing, storing, and querying biomedical terminologies from simple to sublime. *J Am Med Inform Assoc* 2009;16:305–15.

45 CTS2. http://informatics.mayo.edu/cts2/index.php/Main_Page (accessed 14 Jun 2011).

46 LexEVS 6.0 CTS2 Guide—EVS—LexEVS—National Cancer Institute—Confluence Wiki. https://wiki.nci.nih.gov/display/LexEVS/LexEVS+6.0+CTS2+Guide (accessed 2 Jul 2012).

47 caDSR and ISO 11179—caDSR—National Cancer Institute—Confluence Wiki. https://wiki.nci.nih.gov/display/caDSR/caDSR+and+ISO+11179 (accessed 8 Dec 2012).

48 Warzel DB, Andonyadis C, McCurry B, *et al*. Common Data Element (CDE) management and deployment in clinical trials. *AMIA Annu Symp Proc* 2003;2003:1048.

49 Terminfo Project—Overview. http://www.hl7.org/Special/committees/terminfo/overview.cfm (accessed 9 Dec 2012).

50 NIVEL|LINH. http://www.nivel.nl/en/netherlands-information-network-general-practice-jlinh (accessed 28 Jul 2012).

51 NIVEL|Netherlands institute for health services research. http://www.nivel.nl/en (accessed 11 Apr 2012).

52 Clinical Practice Research Datalink—CPRD. http://www.cprd.com/intro.asp (accessed 28 Jul 2012).

53 Medicines and Healthcare products Regulatory Agency. http://www.mhra.gov.uk/#page=DynamicListMedicines (accessed 28 Jul 2012).

54 Lenzerini M. Data integration: a theoretical perspective. In: *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems; 3–6 June 2002*, New York, NY: ACM, 2002:233–46.

55 Smith B, Ceusters W. Ontological realism: a methodology for coordinated evolution of scientific ontologies. *Appl Ontol* 2010;5:139–88.

56 Grenon P, Smith B. SNAP and SPAN: Towards dynamic spatial ontology. *Spat Cogn Comput* 2004;4:69–104.

57 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32(database issue):D267–70.

58 Scheuermann RH, Ceusters W, Smith B. Toward an ontological treatment of disease and diagnosis. *AMIA Summit on Translational Bioinformatics* 2009;2009:116–20.

59 Goldfain A, Smith B, Arabandi S, *et al*. Vital sign ontology. In: *Proceedings of the Workshop on Bio-Ontologies*, Vienna, ISMB, 2011:71–4.

60 information-artifact-ontology—The Information Artifact Ontology (IAO) is an ontology of information entities based on the BFO—Google Project Hosting. http://code.google.com/p/information-artifact-ontology/ (accessed 9 Dec 2012).

61 Brinkman RR, Courtot M, Derom D, *et al*. Modeling biomedical experimental processes with OBI. *J Biomed Semantics* 2010;1(Suppl. 1):S7.

62 Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25:25–9.

63 Luciano JS, Andersson B, Batchelor C, *et al*. The translational medicine ontology and knowledge base: driving personalized medicine by bridging the gap between bench and bedside. *J Biomed Semantics* 2011;2(Suppl. 2):S1.

64 Leysen P, Bastiaens H, Van Royen P. TRANSFoRm : development of use cases. http://transformproject.eu/Deliverable_List_files/D1.1%20Detailed%20Use%20Cases_V2.1-2.pdf (accessed 28 Feb 2013).

65 Qamar R, Rector A. Semantic issues in integrating data from different models to achieve data interoperability. *Stud Health Technol Inform* 2007;129:674–8.

66 Park J, Ram S. Information systems interoperability: what lies beneath? *ACM Transactions on Information Systems* 2004;22:595–632.

67 Pathak J, Wang J, Kashyap S, *et al*. Mapping clinical phenotype data elements to standardized metadata repositories and controlled sterminologies: the eMERGE network experience. *J Am Med Inform Assoc* 2011;18:376–86.

68 Shvaiko P, Euzenat J. Ontology matching: state of the art and future challenges. *IEEE Trans Knowl Data Eng* 2013;25:158–76.

69 Choi N, Song I-Y, Han H. A survey on ontology mapping. *ACM SIGMOD Rec* 2006;35:34–41.

70 Shvaiko P, Euzenat J. A survey of schema-based matching approaches. In: Spaccapietra S (ed) *Journal on data semantics IV*. Berlin/Heidelberg: Springer, 2005: 146–71.

71 CTS2—HL7Wiki. http://wiki.hl7.org/index.php?title=CTS2 (accessed 28 Jul 2012).

72 Cimino JJ, Clayton PD, Hripcsak G, *et al*. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *J Am Med Inform Assoc* 1994;1:35–50.

73 Cimino JJ. Terminology tools: state of the art and practical lessons. *Methods Inf Med* 2001;40:298–306.

74 Miller GC, Britt H. A new drug classification for computer systems: the ATC extension code. *Int J Biomed Comput* 1995;40:121–4.

75 BNF.org. http://www.bnf.org/bnf/index.htm (accessed 10 Dec 2012).

Implementation in primary care

The approach described in the article published in JAMIA (2013) is one that is generalizable to multiple contexts, focusing primarily on providing a method to support binding of structural and terminological models in context of local-as-view data mediation. The next step was to implement it in the context of primary care to confirm that it could support the required operations specific to a LHS in primary care.

More precisely, the concrete implementation of the general information model needed to be analysed. Named the Clinical Data Integration Model (CDIM), it is expressed as an ontology and its content relating to primary care is described in this Methods of Information in Medicine article published ahead of print in 2014. (Ethier *et al.* 2015)

It is important to recognize that information models and terminological models are really on a continuum. Some very high level concepts like date of birth are more easily attributed to information models, while other extremely granular information like pseudohypoparathyroidism belong more naturally to terminologies. But concepts like diabetes are in the middle. Design choices applied to guide the creation of CDIM are presented in the article.

The set of the requirements for CIDM to be used in primary care are presented in terms of clinical, research but also pragmatic aspects. Being expressed as an ontology, core design choices are also discussed including the realist versus cognitive approach to ontology development, or the importation of existing realist ontologies to build the intermediate level. (Grenon 2003; Grenon *et al.* 2004)

Finally, advantages like support for provenance (audit) operations or reusability are contrasted with limitations of the approach like the lower control over content and definitions of concepts given the use of terminologies which are outside the control of TRANSFoRm. In the end, the article demonstrates that CDIM is flexible, modular and that it can support the semantic expressions required by a primary care LHS.

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 2014; 53: ■–■
doi: 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

[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.

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 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

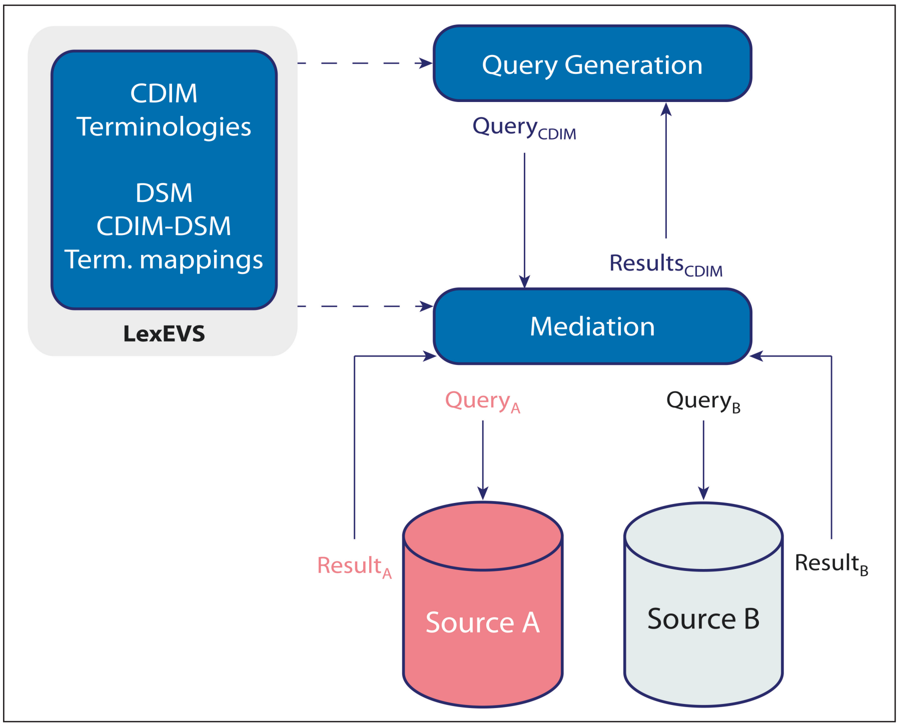


Figure 1 Interoperability framework based on CDIM in context of the Query Formulation Workbench. CDIM and terminologies are bound together to express queries independently of specific sources. Data source models (DSM), CDIM to DSM mappings and terminology mappings are used to translate the query during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

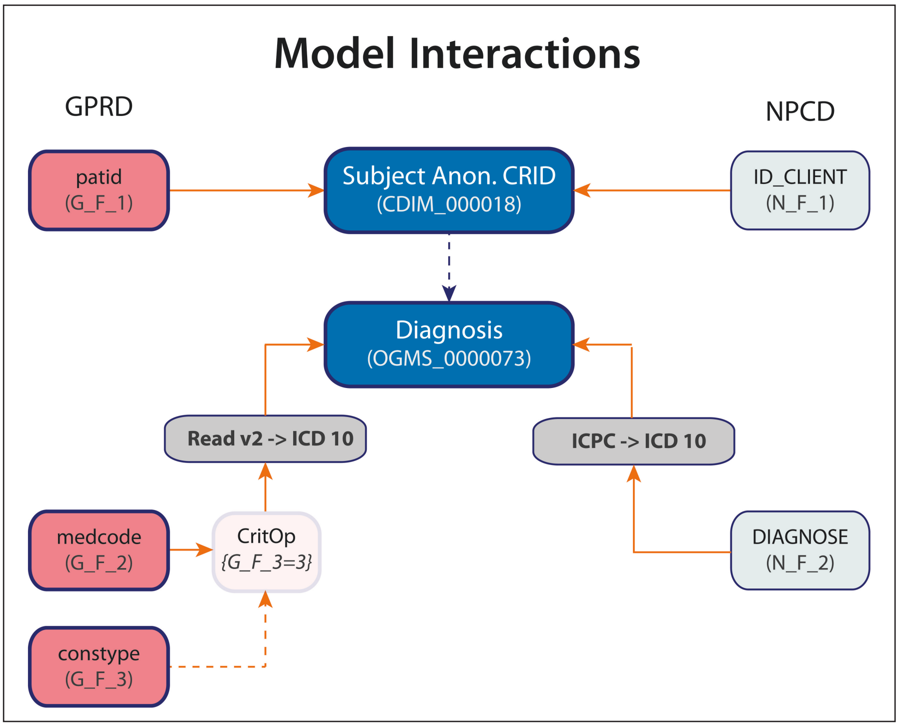


Figure 2 Model interactions in the task of retrieving a list of patient identifiers and diagnoses. In the GPRD database, the "medcode" field contains a diagnosis only if the field "constype" is equal to 3 for the same record. (DSM: GPRD and NPCD; blue boxes – CDIM classes, grey boxes – terminological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

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.

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.

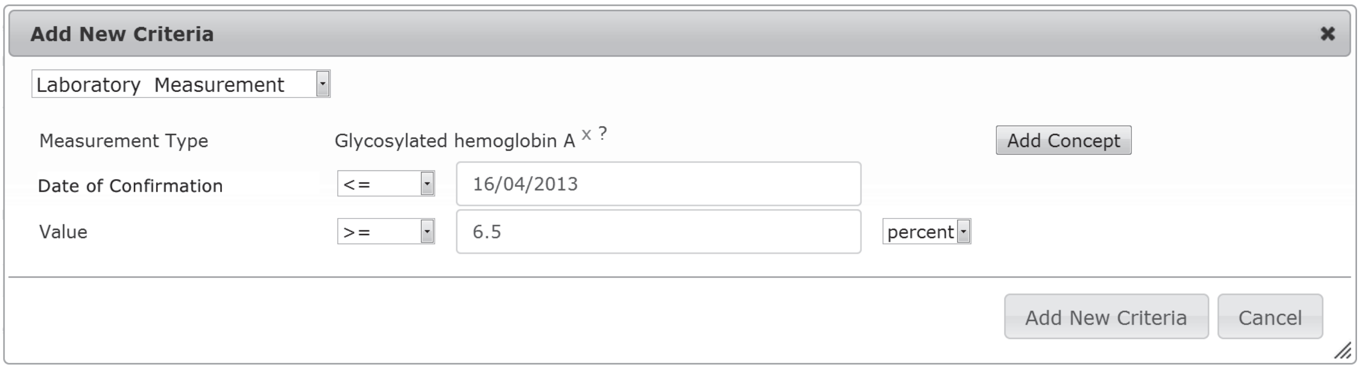
Specific classes, such as *reason for health care encounter* or *health care episode*, were designed in such a way as to avoid inconsistencies with other common classes. For example, a *reason for health care encounter* was formalized as an entity bearing a special role that we called the *reason for 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

[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]



Add New Criteria ×

Laboratory Measurement

Measurement Type: Glycosylated hemoglobin A × ?

Date of Confirmation: <= 16/04/2013

Value: >= 6.5 percent

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.

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 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.

Exemplar results

This fourth article submitted to JAMIA in 2016 presents an exemplar application of the LHS and how it can leverage synergies between its components, here clinical and research activities. It describes how a LHS can facilitate and optimise conduct of prospective research in a primary care practice setting.

Clinical trials have been in a long-term crisis, in terms of rising costs and poor recruitment, at least for the best part of a decade. (Reynolds 2011) However, increasing requirements for safety and effectiveness in routine healthcare demand efficient conduct of clinical trials in ‘real world’ settings. One solution to this problem is to reduce costs for clinical trials by recruiting with prompts in the electronic health record (EHR) system, at the same time doing this in routine clinical practice.

Three of the most difficult tasks, when running trials in primary care in particular, are the identification of suitable research sites, recruitment of patients and the collection of the clinical outcome measures (through research forms). The unified interoperability framework implemented using CDIM as the conceptual model can help with these tasks.

Patient data from EHR can be extracted and presented to the platform as an extensible markup language (XML) file. No specific export structure is forced on the EHR vendors. They all preferred to use XML but each used a different structure for their extraction. The XML documents containing patient data are integrated as data sources and are described using a local model (one per type of XML structure, in this case one per vendor) to participate in the mediation platform. The local models are then mapped to CDIM to enable correct semantic use.

Once in place, various research activities can benefit from this approach. Feasibility studies can be run by executing the inclusion and exclusion criteria against each practice to identify sites with the most promising populations. A tool can be included in the EHR to identify potentially eligible patients for a study. With each new patient visit, patient information is extracted and inclusion/exclusion criteria are executed against his data. If an exclusion is raised, the patient is deemed not eligible and nothing will happen. If no exclusion is triggered, the patient is potentially eligible and a graphical notification (varying per EHR vendor’s choice) is turned on in the EHR to alert the physician. If the physician wishes to include the patient, an electronic case report form is created by the research platform, available EHR data is preloaded in the form and the form is then presented to the physician. (Köpcke *et al.* 2013)

In order to make this approach useful on a larger scale, a bridge to research standards (protocol, forms, data collection) is required. This is the second aspect demonstrated in the article. The Clinical Data Interchange Standards Consortium (CDISC) is an open, multidisciplinary, neutral, non-profit standards developing organization that has been working through productive, consensus-based collaborative teams, since its formation in 1997, to develop global standards and innovations to streamline medical research and ensure a link with healthcare. (CDISC | Strength Through Collaboration) It produces standards used by the research community to guide many processes from protocol descriptions to data collection. The key link is the Operational Data Model (ODM) which contains criteria and form structural elements, as well as data from these forms once filled in. (Bruland *et al.* 2012) By linking ODM with CDIM and the TRANSFoRm platform, a protocol can be drafted with ODM, distributed to multiple primary care research sites using different EHR products and be used locally transparently. The link between ODM and CDIM is what enables the automated execution of criteria on EHR data to identify patients and a transfer of information between the EHR and the research form to preload available information.

As a result, a prospective randomized control trial can be developed using CDISC research standards, linked to queries compatible with the TRANSFoRm LHS and then executed in multiple practices simultaneously.

Title:

Implementation and validation of a standards-based approach to embedding clinical trial functionality in routine electronic health record systems

Corresponding Author:

Name: Jean-Francois Ethier
Address: Centre de Recherche des Cordeliers 15 rue de l'École de Médecine, 75006 PARIS
e-mail: ethierj@gmail.com
Telephone: (+1) 819-346-1110

Authors:

1) Jean-Francois Ethier
INSERM 1138, eq 22
Université Paris-Descartes
Paris, France

2) Brendan C Delaney
Department of Surgery and Cancer
Imperial College London
London, United Kingdom

3) Vasa Curcin
Department of Informatics
King’s College London
London, United Kingdom

4) Mark M. McGilchrist
Public Health Sciences
University of Dundee
Dundee, United Kingdom

5) Sarah N. Lim Choi Keung
Institute of Digital Healthcare, WMG
University of Warwick
Warwick, United Kingdom

6) Lei Zhao
Institute of Digital Healthcare, WMG
University of Warwick
Warwick, United Kingdom

7) Anna Andreasson
Department of General Practice
Karolinska Institute
Stockholm, Sweden

8) Piotr Bródka
Department of Computational Intelligence
Wroclaw Institute of Technology
Wroclaw, Poland

9) Michalski Radoslaw
Department of Computational Intelligence
Wroclaw Institute of Technology
Wroclaw, Poland

10) Nikolaos Mastellos
Department of Primary Care and Public Health
Imperial College London
London, United Kingdom

11) Anita Burgun
INSERM 1138, eq 22
Université Paris-Descartes
Paris, France

12) Theodoros N. Arvanitis
Institute of Digital Healthcare, WMG
University of Warwick
Warwick, United Kingdom

Keywords: Electronic Health Records, Clinical Trial, Learning Health System, Interoperability, Operational Data Model

Word count: 4668

ABSTRACT

Objective: The Learning Health System (LHS) requires integration of research into routine practice. 'eSource' or embedding clinical trial functionalities into routine electronic health record (EHR) systems has long been put forward as a solution to the rising costs of research. We aimed to create and validate an eSource solution that would be readily extensible as part of a LHS.

Materials and Methods: The EU FP7 TRANSFoRm project's approach is based on dual modeling, using the Clinical Research Information Model (CRIM) and the Clinical Data Integration Model of meaning (CDIM) to bridge the gap between clinical and research data structures, using the CDISC Operational Data Model (ODM) standard. Validation against GCP requirements was conducted in a clinical site.

Results: Using the form definition element of ODM, we linked precisely modelled data queries to data elements, constrained against CDIM concepts, to enable automated patient identification for specific protocols and pre-population of electronic case report forms (e-CRF).

Discussion: While initiatives such as IHE profiles, and ISO11179 metadata repositories have been developed, these are too complex for implementation in a low-resource, heterogeneous, highly distributed environment, implied by a scalable LHS. The TRANSFoRm approach provides an ontologically-based alternative that allows precise prospective mapping of data elements in the EHR.

Conclusion: We demonstrated that leveraging the EHR to identify patients and pre-populate e-CRF can be done using a standards-based, reusable approach based on a core information model (CRIM) and a model of meaning (CDIM) to bridge the gap between the clinical and research worlds.

INTRODUCTION

Clinical trials have been in a long-term crisis, in terms of rising costs and poor recruitment, for the best part of a decade.[1] However, increasing requirements for evidence of safety and effectiveness in routine healthcare demand efficient conduct of clinical trials in 'real world' settings.[2] One solution to this problem is to reduce costs for clinical trials by recruiting with prompts in the electronic health record (EHR) system during routine clinical practice.[3] The concept of the Learning Health System (LHS) takes this a step further and envisages a health care organisation where routine EHR systems are the direct mediators of both research and knowledge translation activities.[4] Although scope exists for much of this learning to take place by the analysis of routine data, formal randomised controlled trials will always be required for regulatory purposes, where risks and benefits are finely balanced or where data has been conflicting.[5]

In order to complete successfully a randomised controlled trial (RCT) it is necessary to: (1) establish the feasibility of the study eligibility criteria; (2) identify suitable sites; (3) identify suitable subjects; (4) obtain consent and collect baseline data; (5) randomise to intervention and alternative; (6) collect clinical follow up data; (7) collect patient related outcome data; and, (8) transfer data for analysis.[6] Three of the most difficult tasks, when running trials in primary care in particular, are the identification of sites and subjects and the collection of the clinical outcome measures.[7,8] It is proposed that a LHS, embedding research into routine EHR systems, could automate a substantial part of the trial's screening process.[9] Namely, eligibility criteria can be tested against EHR patient data. Patients meeting at least one exclusion criterion will not be processed further, thereby saving substantial resources by not incurring manual review. Secondly, for patients potentially eligible, the eligibility form can be pre-filled with data present in the EHR, in order to minimise unnecessary manual entry. A similar process can be used to pre-populate electronic case report forms (eCRFs). In addition, clinical data collected within a trial should be made available in the EHR, in order to enhance routine clinical care and safety monitoring.[10] The ability to place trial information in routine EHRs, at the point of collection, would be a significant step towards safer and more efficient clinical trials.[11]

There has been a steady move away from paper case report forms (CRFs) towards electronic data capture (EDC) systems, enabling direct collection of data into digital form, referred to as eSource.[12] Good Clinical Practice principles need to be adopted to ensure that the requisite standards are in place for eSource, while changes are made to the data collection process and governing regulations to fit in with this electronic context. There are three models of eSource currently being explored: 1) entry into an EDC system with copying to the EHR, 2) entry into a 3rd party system with copying to both the EHR and the EDC and 3) collection within the EHR with copying to the EDC. Local preferences, maturity of EHR systems and sponsor requirements are likely to maintain this heterogeneous approach, emphasising the paramount importance of adherence to standards.

There are two important standards bodies in this area, The Clinical Data Interchange Standards Consortium (CDISC) and Health Level Seven (HL7) International. Firstly, CDISC have established a suite of standards for Clinical Trials. (www.cdisc.org/foundation-standards). These include specification of a trial protocol (Protocol Representation Model/ PRM) and its data model representing a CRF (Operational data Model/ODM), specification of study design (Study Design Model/SDM), specification of tabulated data (Study Data Tabulation Model/ SDTM) and standardised sets of defined data elements for use in the above (Clinical Data Acquisition Data Standards Harmonisation/CDASH). Semantic resources also include CDISC controlled terminology sets, maintained and distributed as part of the National Cancer Institute's (NCI) Thesaurus, available within the NCI's Enterprise Vocabulary Service; these terminology sets are also available via the CDISC eSHARE. Secondly, within the healthcare domain, HL7 provides interoperability standards, with the HL7 Reference

Information Model (RIM) and, more recently, the HL7 Fast Healthcare Interoperability Resources Specification (FHIR) being models of structure for health data capture and exchange. [http://www.hl7.org/implement/standards/rim.cfm and http://www.hl7.org/implement/standards/fhir/, respectively] The Biomedical Research Domain Analysis Model (BRIDG) provides a high level view of clinical research activities with links to HL7 RIM concepts. (http://www.bridgmodel.org) The Primary Care Research Object Model v 3 (PCROM v3) is a domain specific model for primary care trials that maps to BRIDG v3.[13]

To enable a LHS for clinical trials, it is first necessary to consider how existing standards around EDC can be deployed as 'e-Source'. There are parallels between the research requirements of Good Clinical Practice (GCP) and the needs of an EHR system, in terms of contemporaneous, complete, accurate records with audit of changes (controlled data) and secure, safe storage.[12]

CDISC Healthcare Link (HCL) and Integrating the Healthcare Enterprise (IHE)

There have been several efforts to go beyond simple semantic interoperability, in order to enable CDISC standards to be used to support EHR systems as eSource as part of the CDISC HCL initiative. The principal mover in this has been IHE (www.ihe.org), through a collaboration representing the clinical trials community (e.g., Contract Research Organisations CROs, Biotechnology & Pharmaceutical industry) and EHR vendors.[14] IHE has developed a set of profiles for this purpose. The two relevant profiles are Retrieve Form for Data Capture (RFD) and Retrieve Process for Execution (RPE). An additional content profile, the Clinical Research Document (CRD), enables data to be extracted from an EHR system using HL7's Continuity of Care Document (CCD) specification, with data elements specified via CDASH and used to pre-populate an eCRF provided by a specified form generator. The RPE profile specifies when in workflow the forms should be completed.

Several proof-of-concept studies using IHE profiles have been completed. These include, integration of Common Data Elements from NCBI CaBIG EVS into the RFD profile,[15] and STARBRITE, a single site proof of concept implementation within a heart failure clinical trial.[16] STARBRITE showed that, in the context of two live patient encounters, valid data could be obtained from the EHR; however, significant challenges in precise semantic definitions of extracted information remained. Strikingly, the fact that in this setting, a cardiology clinic, the CRF was more precisely defined and contained more finely grained information, than a contemporaneous clinic note, led to the suggestion that the CRF should be the primary source and the EHR a subset of this information.

TRANSFoRm Project

The European FP7 TRANSFoRm project aimed at developing an infrastructure for a Learning Health System in European Primary Care (www.transformproject.eu). Primary Care represents the ultimate low-resource, heterogeneous, highly distributed environment, especially when the multi-language, multi-health system dimensions of Europe are added. As part of this vision, we have developed a modular, model-based system for conducting clinical trials and implemented it within five different EHR systems in four European member states. TRANSFoRm studied the progress of IHE during 2012-13 and examined its capability alongside the requirements of a TRANSFoRm use-case and proposed evaluation study.[17] There would be significant challenges for TRANSFoRm in implementing the current IHE approach. Only single EHR systems had been harnessed to deploy standard forms and pre-populate them with limited data from the EHR. In each case, the data collection requirements of the study shared, or at least conformed to a minimum set of data elements defined in CDASH, and that the EHR system was capable of supporting the technical infrastructure required to operate the CRFs. The requirements implied a large academic centre for conducting trials and with support for a complex IT infrastructure. The TRANSFoRm unified interoperability framework takes into account the necessity for structural and terminological models to be bound together in order to derive the full semantic meaning of data.[4,18] It is based on the mediation method with a local-as-view paradigm. This is especially important since TRANSFoRm does not control the source of data, the EHR, and cannot force a common structure to support data federation. TRANSFoRm therefore had to consider the requirements of multi-site, multi-system data collection with a low resource overhead, such that the LHS can encompass all of a healthcare system, not just large academic centres. In addition, existing limits on the coverage of CDASH and the need to add new data elements has limited the adoption of eSource to demonstrator projects. What is required is a readily extensible framework, enabling researchers to define clinical data elements to research standards like CDISC and to semantically align them on native EHR systems data. TRANSFoRm has taken an approach of using existing CDISC standards, but referencing a core data model, expressed as an ontology, to provide a flexible and more streamlined approach.

In this paper, we describe TRANSFoRm's approach to embedding clinical trial functionality within the EHR systems and enabling the pre-population of eCRFs directly from the EHRs, where the data is available, and the potential of recording of CRF data, collected during research, within the EHR system. This paper describes the methods we adopted to create semantically enriched and model-based extensions to existing standards, how we implement these in live EHR systems, and how we validated the technical functionality of the approach.

MATERIAL AND METHOD

General approach

TRANSFoRm's dual-level modelling approach separates the stable domain information from the heterogeneous data sources[19] to achieve structural and semantic interoperability between different actors in the LHS (clinical investigators, EHRs, researchers, CRFs).[18–20] The first level defines the requirements and workflow of the research process. In our implementation, this role is fulfilled by the Clinical Research Information Model

(CRIM),[13] a domain-specific implementation of BRIDG.[21] So, it is the role of CRIM to establish data elements in research workflows, e.g. query structures for retrieving patient study data from aggregated EHR data repositories. At the second level, the clinical primary care domain is specified, using the Clinical Data Integration Model (CDIM) ontology,[22] that fleshes out CRIM concepts with well-defined clinical concepts.

CDIM offers a unified view of the primary care domain and enables users to work with data and express queries using neutral clinical concepts, without needing any knowledge about the specific schema of the target data source. CDIM has been developed as a realist ontology, all of whose classes have instances in the real world, and works in conjunction with medical terminologies that provide concrete instantiations. CRIM and CDIM models together specify the data flow through TRANSFoRm's LHS infrastructure.

ODM and Dual Modelling

The CDISC Operational Data Model (ODM) is a vendor neutral, platform-independent XML format for interchange and archive of clinical study data, designed to facilitate regulatory-compliant acquisition, archive and interchange of data and metadata for clinical research studies. (http://www.cdisc.org/odm). ODM's `<FormDef>` element captures eCRF composition and structure, with `<ItemGroupDef>` elements used to group related items (e.g. a systolic blood pressure measurement value, the unit of measure and the time at which it was measured), with `<ItemDef>` elements containing specific item metadata.

While TRANSFoRm was conceived as a dual modelling approach, with specific requirement of pre-populating forms from the EHR and writing back to the EHR, ODM has not been designed with this in mind. Nevertheless, a natural parallel exists and we used `<ItemGroupDef>` to define the research unique identifier, referencing research data queries expressed using CRIM, and `<ItemDef>` to define clinical data elements referencing the CDIM ontology. As both these elements are contained within <FormDef>, the necessary model constraints are applied.

In order to embed an EHR data extraction request into ODM, <ItemGroupDef> was extended by adding a <QueryId> child element, containing a unique identifier linking the item group with the corresponding query as illustrated in figure 1. CDIM is used to annotate <ItemDef> through its <Alias> element, which allows binding to an external model using the context attribute (e.g. CDIM_2.2) and the value attribute (e.g. CDIM_000070). In this way, TRANSFoRm's dual-level modelling enables data interoperability between EHR patient data and the ODM.

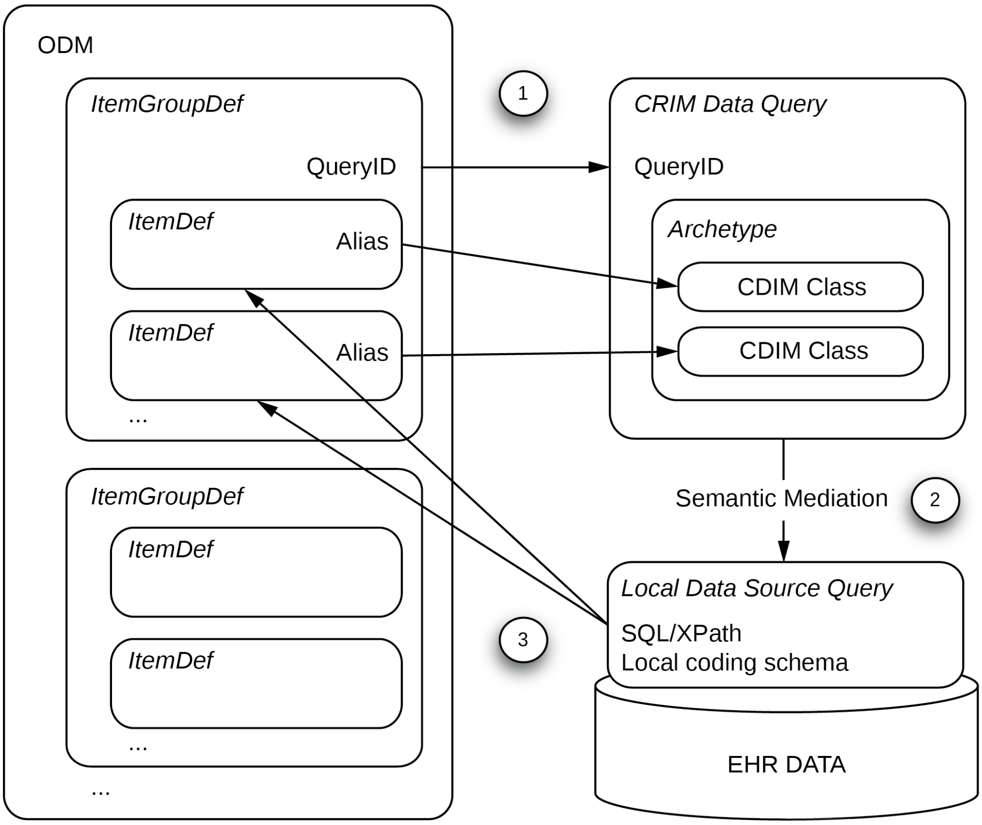


Figure 1 : Models interactions. 1) Link between ODM and the queries. 2) Semantic mediation translating and executing the query on the patient extract. 3) Pre-populating data from the EHR.

The CDIM terms embedded into the CRIM queries are organised into archetypes, computable expressions of a domain content model in the form of structured constraint statements, based on a reference model.[20] The TRANSFoRm archetypes are defined using the openEHR archetype definition language (ADL) and reference CDIM classes to ascertain semantics. While openEHR uses a hierarchical reference model,[23] TRANSFoRm uses a simpler event-based tabular structure, making it more compatible with CDISC data formats and patient data structures present in the current clinical data landscape. Archetypes are then used as part of Data Extraction Query, which also contains the rest of the logic to be applied (e.g. select first blood pressure, or most recent…) and is related to the corresponding `<ItemGroupDef>` through the `<QueryId>` element. This query does not contain source specific structural information and is used for every source. Once the embedded Data Extraction Queries have been translated for the specific source by the TRANSFoRm interoperability framework and executed, the results are annotated with CDIM concepts and placed in the proper `<ItemDef>`, as identified by the `<Alias>` element.

Validation Process

Preparation for the gastro-oesophageal reflux disease (GORD) evaluation study took place between late 2013 and early 2015 in Poland, beginning with the integration of the vendor system (mMedica from Asseco Poland S.A.) and the TRANSFoRm platform through a single platform component, the data node connector (DNC), which brokers communication with the TRANSFoRm study system and other platform components. The intention of the validation study was to ascertain the accurate functioning of the TRANSFoRm tools and to carry out a GCP certification.

A simulated study with 10 process scenarios was designed.[24] The software was installed in practices and data collection scenarios carried out by Polish clinicians and TRANSFoRm staff acting as patients. The training plans were also designed and tested with pilot users. The installation and regular operation of TRANSFoRm components, including data collection tools, TRANSFoRm study system and the data node connector, were documented through a set of Installation Qualification, Operation Qualification and Performance Qualification tests, all performed on the pilot trial site. The development teams involved in software production were themselves assessed in terms of training, software quality assurance procedures and institutional policies.

RESULTS

Implementation of the research platform components and workflows

The overview of the TRANSFoRm implementation approach for embedding clinical research into clinical care is shown in figure 2. The TRANSFoRm Study System (TSS) is centrally hosted at a secure location and holds the study information and protocols, defined via ODM/SDM files. It also acts as the research repository for the collected eCRF data. Data collection form templates are automatically created from ODM definitions, extended with generic query definitions and later instantiated as EHR-specific queries for each clinical site.

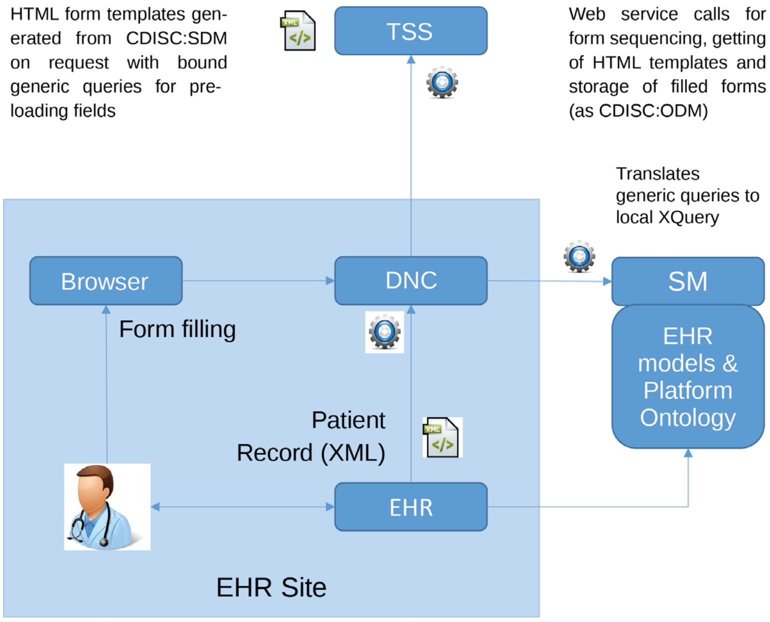

HTML form templates generated from CDISC:SDM on request with bound generic queries for pre-loading fields
TSS
Web service calls for form sequencing, getting of HTML templates and storage of filled forms (as CDISC:ODM)
Translates generic queries to local XQuery
Browser
DNC
SM
Form filling
EHR models & Platform Ontology
Patient Record (XML)
EHR
EHR Site

Figure 2: TRANSFoRm project research platform

The coordination of study activities at the local level is performed by the data node connector (DNC) components, which sit locally to the EHR instances, or a single instance in case of a centrally hosted EHR. All research data capture operations (e.g. eligibility checking), eCRF pre-population and manual eCRF completion, are orchestrated and performed by the DNC. Thus the data flow between the EHR and the DNC remains local to the EHR and only data identified for research purposes is sent to the research repository in line with the project's security and data protection framework.[25] The TSS cannot issue commands to any DNC, only the DNC can initiate communication with the TSS. The DNC can pull data from the TSS but the TSS cannot push data to nor pull data from the DNC as initiator of the communication. The latter is important as often the DNC will sit behind a clinical organisational firewall.

The DNC is started with the host EHR system and obtains from the TSS information about the currently active study protocols and their eligibility criteria. The generic study definitions are then translated by the Semantic Mediator (SM) component into locally executable queries. When a patient arrives for a consultation, his record is sent to the DNC as an XML document, where it is checked for eligibility. The TRANSFoRm platform does not mandate a specific structure or model for that file and indeed, many different file structures are used in the project. While the TRANSFoRm platform could also support other types of data structure such as relational databases, the XML patient record extract was the preferred solution by the EHR vendors.

When a presenting patient is found to be potentially eligible for a study, the DNC notifies the clinician of the eligibility via a pop-up message requesting completion of eligibility checks and consent/randomisation. Thereafter when the recruited patient presents at the practice, the DNC retrieves the appropriate eCRF forms from the study system, transported as HTML forms parameterised for pre-loading and storage of field values, together with the corresponding CDISC ODM document container with the ClinicalData section parameterised to store the data values entered into HTML fields. The generation and parameterisation of the HTML and ODM documents is performed by the TSS based on a pre-established ODM to interface translation, OID, QueryID and CDIM alias. Using this information, the DNC can correctly pre-load form fields by applying the queries to the patient data extract and inserting the resulting values at the corresponding place in the form. The pre-loaded HTML form can then be presented to the clinician for validation and entry of data items that have not been pre-loaded. The form can either be embedded into the EHR or accessed through a web browser.

Once approved, the form is submitted to the DNC where data is inserted in the ODM file. The DNC then sends the ODM document containing the responses from the form to the TSS for research activities. It also sends the associated ODM files and HTLM forms to the EHR for auditing purposes and reviewing by clinicians at a later date, if required. The EHRs currently partnering with TRANSFoRm do not store the form field data as individually coded facts, but as a single artefact which can be viewed as a whole. While the mappings between the patient data extract and CDIM could potentially be used to support granular transfers back to the EHR, the vendors preferred not to explore this aspect in the first version.

While the EHR triggers the DNC when a new patient arrives for consultation, the tasks of presentation, filling and acquisition of form data are all orchestrated by the DNC.

Data Flow and Document Examples

The study ODM document contains the form details that are relevant to the study data collection. A section of the ODM document describing a question about a history of GORD for a patient is presented in figure 3. Of note, the ODM includes multilingual support as required in TRANSFoRm.

```
<ItemGroupDef OID="IG.GORDDX" Name="GORD Diagnosis ItemGroup" Repeating="No">
        <Description>
                <TranslatedText xml:lang="en">GORD diagnosis</TranslatedText>
                <TranslatedText xml:lang="nl">Diagnose GORZ (gastro-oesofagale refluxziekte)</TranslatedText>
                <TranslatedText xml:lang="pl">Rozpoznanie choroby refluksowej przełyku</TranslatedText>
        </Description>
        <ItemRef OrderNumber="1" ItemOID="ID.GORDDX" Mandatory="Yes"/>
        <ItemRef OrderNumber="2" ItemOID="ID.GORDDX_DATE" Mandatory="No"/>
        <transform:QueryId>2a14cf7e-4d0c-4076-983c-d5a4ab3cb7be</transform:QueryId>
</ItemGroupDef>

<ItemDef OID="ID.GORDDX" Name="GORD Dx Item" DataType="text" Length="3">
        <Question>
                <TranslatedText xml:lang="en">Have you been previously diagnosed with GORD?</TranslatedText>
                <TranslatedText xml:lang="nl">Is bij u eerder de diagnose GORZ gesteld?</TranslatedText>
                <TranslatedText xml:lang="pl">Czy zdiagnozowano, wcześniej, u Pana/Pani chorobę refluksową przełyku. </TranslatedText>
        </Question>
        <CodeListRef CodeListOID="CL.YN"/>
        <Alias context="CDIM_2_2">OGMS_0000073</Alias>
</ItemDef>

<ItemDef OID="ID.GORDDX_DATE" Name="GORD Dx Date Item" DataType="date" Length="10">
        <Question>
                <TranslatedText xml:lang="en">Date of latest GORD diagnosis</TranslatedText>
                <TranslatedText xml:lang="nl">Datum van eerdere diagnose GORZ</TranslatedText>
                <TranslatedText xml:lang="pl">Data wcześniejszej diagnozy choroby refluksowej przełyku.</TranslatedText>
        </Question>
        <Alias context="CDIM_2_2">CDIM_000012</Alias>
</ItemDef>
```

Figure 3: ODM example

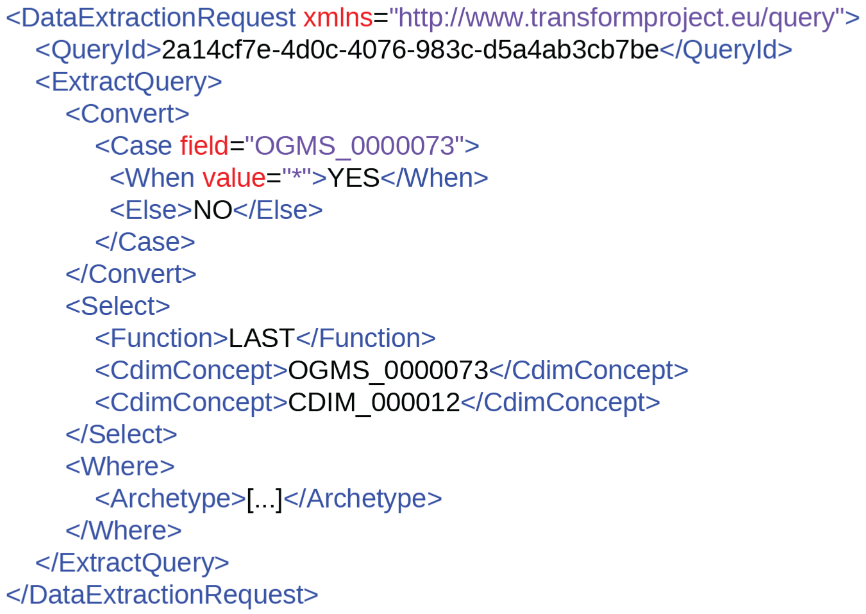
The ODM form elements are translated into HTML form elements suitable for web or mobile devices using standardised templates. The English HTML form of the first ODM item in figure 3 is presented in figure 4.

```
<span class="itemtitle">Have you been previously diagnosed with GORD?:</span>
<select id="field0" name="IG.GORDDX$ID.GORDDX$value">
        <option value="YES">Yes</option>
        <option value="NO">No</option>
</select
<input id="hiddenInput3" type="hidden" name="IG.GORDDX$ID.GORDDX$preLoadValue"
value="{2a14cf7e-4d0c-4076-983c-d5a4ab3cb7be$CDIM_0000073}" />
<script>
        document.getElementById('field0').value = document.getElementById('hiddenInput3').value;
</script>
```

Figure 4 : HTML form example

When the DNC receives the ODM and HTML form, it uses the related data extraction query to start the pre-population process. The 'hidden input' element is used to carry the anchor for pre-loading while not disturbing the user experience. As illustrated in figure 4, the element contains the associated QueryId value as well as the target CDIM concept. Prior to display, the DNC will insert the pre-loaded value using the hidden field and at display time the script will load the value from the hidden field into the visible field. The hidden field also carries a trace of the value used during pre-loading even if the visible text field value is changed by the user. This is extremely important for audit purposes. A similar approach is used to generate other types of HTML elements like textboxes.

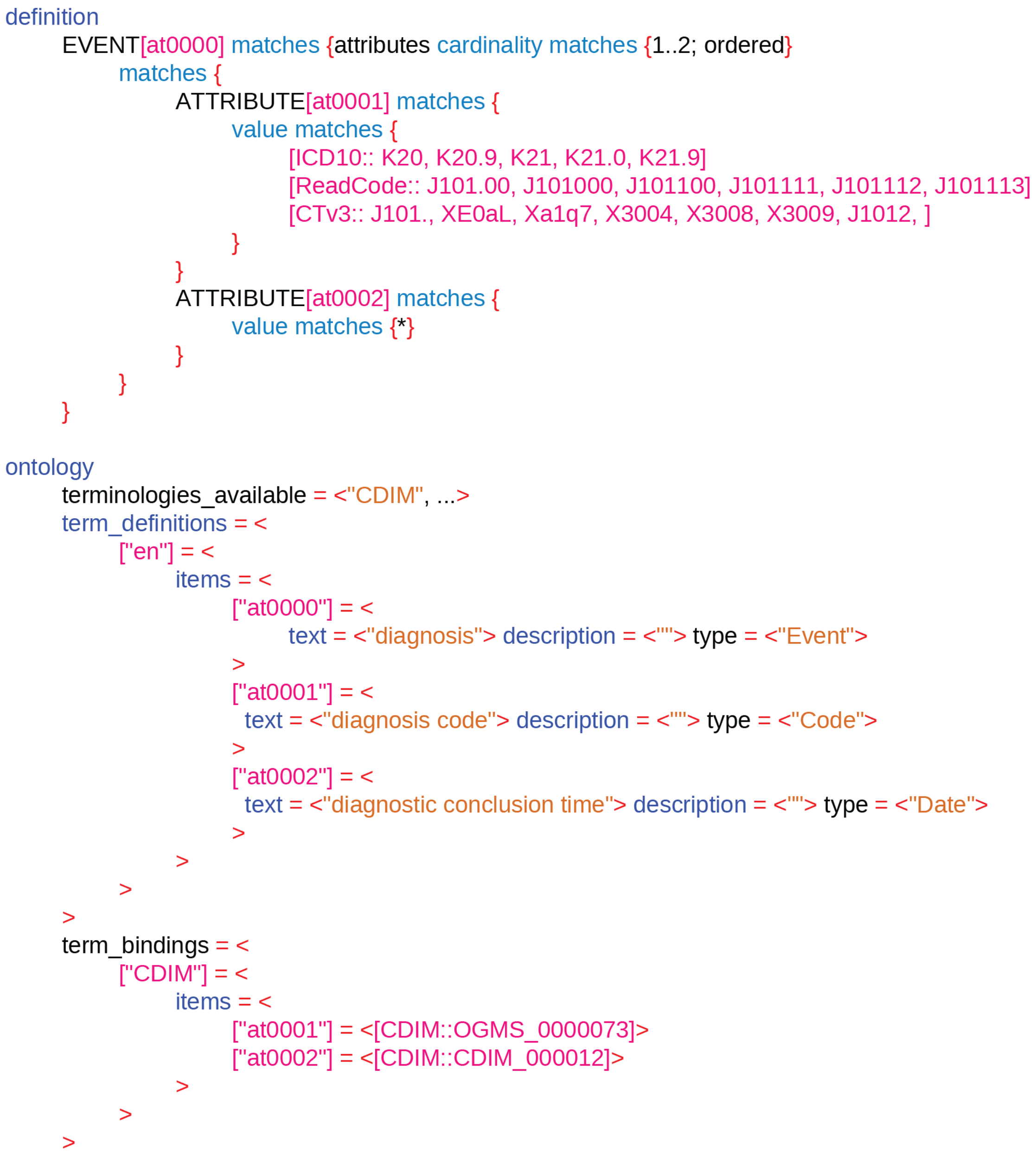
The QueryId is a reference to a Data Extraction Query. Temporal functions and other conversion functions are hosted there, such as the LAST function illustrated in figure 5 representing the GORD diagnosis example, as well as one or more target CDIM concepts.



```
<DataExtractionRequest xmlns="http://www.transformproject.eu/query">
    <QueryId>2a14cf7e-4d0c-4076-983c-d5a4ab3cb7be</QueryId>
    <ExtractQuery>
        <Convert>
            <Case field="OGMS_0000073">
              <When value="*">YES</When>
              <Else>NO</Else>
            </Case>
        </Convert>
        <Select>
            <Function>LAST</Function>
            <CdimConcept>OGMS_0000073</CdimConcept>
            <CdimConcept>CDIM_000012</CdimConcept>
        </Select>
        <Where>
            <Archetype>[...]</Archetype>
        </Where>
    </ExtractQuery>
</DataExtractionRequest>
```

Figure 5: Data Extraction Request example (archetype element collapsed). OGMS_0000073 represents the diagnosis and CDIM_000012 the moment at which the diagnosis was established.

Based on the list of diagnoses fetched from the patient’s data, it will take the last entry. Then, before transferring data, if the data element annotated with OGMS_0000073 has content, it will return the value “YES” otherwise, it will return “NO”. The archetype code has been collapsed in figure 5 to facilitate reading but is presented in figure 6.



```
definition
    EVENT[at0000] matches {attributes cardinality matches {1..2; ordered}
        matches {
            ATTRIBUTE[at0001] matches {
                value matches {
                    [ICD10:: K20, K20.9, K21, K21.0, K21.9]
                    [ReadCode:: J101.00, J101000, J101100, J101111, J101112, J101113]
                    [CTv3:: J101., XE0aL, Xa1q7, X3004, X3008, X3009, J1012, ]
                }
            }
            ATTRIBUTE[at0002] matches {
                value matches {*}
            }
        }
    }

ontology
    terminologies_available = <"CDIM", ...>
    term_definitions = <
        ["en"] = <
            items = <
                ["at0000"] = <
                    text = <"diagnosis"> description = <""> type = <"Event">
                >
                ["at0001"] = <
                  text = <"diagnosis code"> description = <""> type = <"Code">
                >
                ["at0002"] = <
                  text = <"diagnostic conclusion time"> description = <""> type = <"Date">
                >
            >
        >
    >
    term_bindings = <
        ["CDIM"] = <
            items = <
                ["at0001"] = <[CDIM::OGMS_0000073]>
                ["at0002"] = <[CDIM::CDIM_000012]>
            >
        >
    >
```

Figure 6: Archetype extract

The archetype defines the required elements for this data operation (e.g. diagnosis related information), CDIM annotations and constraints. In this example, we are looking specifically for a patient diagnosis (OGMS_0000073) of GORD (as expressed by the chosen terminology codes). To express a diagnosis of GORD in a family member would use another CDIM code but the same terminology codes.

CDIM, in conjunction with relevant terminologies, allows expression of precise, complete and fine-grained clinical concepts as required by a LHS (answering the needs of users as well as catering to data sources with varying granularity). This simple method is anchored into a solid model (CDIM) and allows for a high level of granularity and precision when taking into account the various operators as well as the significant coverage of terminologies, including the UMLS.

Validation study

In order to achieve Good Clinical Practice (GCP) certification, the TRANSFoRm system underwent a series of tests, including Installation Qualification (IQ), Operational Qualification (OQ) and Performance Qualification (PQ). These tests established the functional correctness of the system, with respect to the requirements and specification, and also establishing the integrity of the data that is output from the system. The non-functional aspects examined included training materials, technical support, skill level of the development teams, and software quality assurance procedures.

Post-installation support was provided by members of the TRANSFoRm and vendor teams covering the use of the updated vendor software and supporting TRANSFoRm software. Most issues arose from the sequencing of forms to be filled within the EHR system. In the cases where this was implemented by the vendor, we encountered forms submitted in an invalid order. Extensive logging by the TRANSFoRm DNC meant that a full record of these issues could be maintained and there was less reliance on GP reporting to understand these issues.

DISCUSSION

TRANSFoRm has shown that the process of integrating clinical trial process and data management into the EHR system can both be based on CDISC standards and be largely accomplished by a separate research system (the TSS), minimising the workload on EHR vendors. This is important in lowering the barriers to adoption and increasing the uptake of the LHS, whilst still using current data standards. All a vendor has to do to use our approach is to produce a data source model describing its XML patient data schema and its mapping to the CDIM, as well as providing an API, or equivalent, for communication with the DNC.

We did not use the full ISO 13606 archetype model, which is more complex than our needs, but an archetype-based approach. The potential drawback of this is that, although ADL can be used to construct the archetypes, the lack of a formal dual-ontology means that OWL-based and similar techniques cannot be used to validate these archetypes.[26] Nevertheless, given the formal ontology nature of CDIM, relationships between the concepts used in the archetype can be ascertained. At present TRANSFoRm lacks a simple expression builder, as exists for ISO13606, for researchers to author ODM and SDM compliant XML containing bound data elements with appropriate ontology references (via `<Alias>`). This would not be difficult to develop but lies outside of the scope of the project.

The approach we outlined is generalizable to other domains as different domain specific models can be created. We used higher level concepts to ensure future compatibility by using BRIDG for CRIM and building CDIM from middle level ontologies.[27,28] Standards, like the ISO11179 for metadata registries, were designed to allow the creation of various data elements with different level of granularity and detail. While this flexibility is essential to define data elements purely for research purposes, when applied to clinical interoperability, this can quickly lead to a profusion of slightly varying data elements. This is especially problematic when put in context of organizing data flows between the research data structures and the EHRs. Mappings need to be created between EHRs and data elements. All these slight variations can confuse and complicate mapping creations and maintenance. Moreover, to work effectively, links between the EHRs and the data elements need to be established prospectively in order to avoid the need to create a new mapping each time a new data element is derived with a slightly modified definition.

The ISO11179 standard has been used to define and store common data elements, most notably in the NCI CaDSR. The latest revision of ISO11179 Ed 3 contains a new registry metamodel, separating the concept layer of a data element from its representation (http://metadata-standards.org/11179/). This applies to both individual data elements and sets of concepts, so for example the concepts 'sex' and 'genetic karyotype' are related as data elements and conceptual domains with separate data element representations. Attempts have been made to directly represent ODM data elements using this new metamodel, serving the same basic requirement that we were attempting to solve with TRANSFoRm. This proved not to be possible as all the required attributes of ODM, and clinical representations found in EHRs could not be represented.[29] By using two alias, one to a metadata repository and one to CDIM, the `<ItemDef>` element can become an explicit link.

We propose a model, based on formal ontology, representing a shared understanding of the domain that will provide precise definitions, in this case CDIM. The adoption of CDIM to represent a shared 'model of meaning' allows the separation of definition from implementation.[18] ODM becomes the key standard, with references binding item definitions to CDIM and embedding the research meaning as a precisely defined query, structured and guided by CRIM. By associating the standards and the models using an archetype-based approach we were able to keep the implementation simple, using a connector, rather than requiring EHR vendors to implement a complex stack of templates and profiles. The project has proven its success in being able to implement this approach across five different systems in four separate countries, all running the same research protocol.

As described earlier, an alternative approach has been taken by IHE/HCL whereby the process of integration is undertaken by EHR vendors. The TRANSFoRm approach should not be seen as a competitor to IHE, but an attempt, in the context of an academically-led project to streamline the process of using CDISC standards. Table 1 compares the TRANSFoRm and HCL/IHE approaches.

Requirement	**HCL**	**TRANSFoRm**
Form specification	ODM	ODM
Research CDE definition	CDASH	CDIM Ontology
Research CDE storage and distribution	CDISC SHARE (ISO11179)	Not implemented but would be ISO11179 MDR
Research CDE mapping to clinical DE	DEX	Archetype. CDIM ontology referenced by ItemDef Alias
Pre-population specification of query	SDM (xpath) via DEX	SDM (xpath) via pre-specified queries referenced by ODM ItemGroupDef QueryID
Pre-population extraction of EHR data	RPE and CRD	Via Data Node Connector and EHR API
Semantic mapping	CDASH – restricted code set	TRANSFoRm Terminology Service (LexEVS) augmented by manual term selection and binding
Display of CRFs	RFD – proforma implemented by EHR system	Via Data Node Connector and EHR API
Data storage from CRFs	Archive	TSS (vie Data Node Connector)
Audit and change control	Archive	Provenance model
Security and authorisation		TSS (inherited from local authorisation)

Table 1: Key differences between HCL and TRANSFoRm. HCL - CDISC Healthcare Link; ODM – Operational Data Model; CDE – Common Data Element; CDASH – Clinical Data Acquisition Standards Harmonization; CDISC SHARE – Shared Health and Clinical Research Electronic Library; MDR – Metadata Repository; DEX – IHE Data Element

Exchange; SDM – Study Design Model; RPE – IHE Retrieve Protocol for Execution; CRD – IHE Clinical Research Document; RFD – IHE Retrieve Form for Data Capture; TSS – TRANSFoRm Study System,

A defining principle of a Learning Health System is that it is universal, and we therefore need to be able both to define meaning at system level via an ontology and also enable incorporation of legacy systems where there is not the resource to develop and maintain separate infrastructures for research within the clinical system. Both the TRANSFoRm and IHE approaches can co-exist with, for example, CDASH data elements adopting a reference to CDIM (or another ontology). Another easy way to bind both research and clinical definitions is to use the `<ItemDef>` element as a pivot by using two aliases: one to CDASH and one to CDIM. Maintaining clinical meaning across higher levels of abstraction should be seen as desirable in the context of the LHS. The recent completion of SHARE by CDISC, creating a single repository of both CDISC standards and artefacts from forms to data elements, offers potential to develop an integrated solution whereby TRANSFoRm data elements and queries could be made available via SHARE. The approach presented here has been successfully implemented and the next step, ongoing, is to evaluate the performance of the TRANSFoRm approach in a real-world clinical trial against traditional methods.[17] Once the clinical value has been established, further studies will investigate the impact of the LHS paradigm on the throughput of clinical trials in practices and associated financial impact.

AUTHORSHIP STATEMENT

JFE, AB, MM, BD and VC participated in the development of CDIM, the unified framework and its applications. MM and LZ, with the help of SLCK and TNA, directed the development of the infrastructure for the implementation of the DNC. AA and NM developed the use case and participated in the validation study. PB and MR developed the TSS and the tools for the patients. JFE, BD, VC and SLCK helped to draft the manuscript. All authors critically reviewed the manuscript and approved the final version.

ACKNOWLEDGMENTS

We would like to thank Rebecca Kush, Michael Ibara and Sam Hume from CDISC for their insightful comments on drafts of this paper.

COMPETING INTERESTS

None

FUNDING

This work was supported in part by the European Commission—DG INFSO (FP7 247787).

REFERENCES

1 Reynolds T. Clinical trials: can technology solve the problem of low recruitment? *BMJ* 2011;**342**:d3662. doi:10.1136/bmj.d3662

2 Richesson RL, Hammond WE, Nahm M, *et al.* Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J Am Med Inform Assoc* 2013;**20**:e226–31. doi:10.1136/amiajnl-2013-001926

3 Fiore LD, Brophy M, Ferguson RE, *et al.* A point-of-care clinical trial comparing insulin administered using a sliding scale versus a weight-based regimen. *Clin Trials* 2011;**8**:183–95. doi:10.1177/1740774511398368

4 Friedman C, Rubin J, Brown J, *et al.* Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. *J Am Med Inform Assoc* 2015;**22**:43–50. doi:10.1136/amiajnl-2014-002977

5 (IOM) Medicine I of. *Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care*. 2011.

6 Song FJ, Fry-Smith A, Davenport C, *et al.* Identification and assessment of ongoing

trials in health technology assessment reviews. *Heal Technol Assess* 2004;**8**:iii, 1–87. doi:02-28-01 [pii]

7 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**:54–9. doi:Doi 10.1370/Afm.1313

8 Peterson KA Arvanitis TN, Taweel A, Sandberg EA, Speedie S, Richard Hobbs FD. DBC. A model for the electronic support of practice-based research networks. *Ann Fam Med* 2012;**10**:560.

9 Fiore M, Palassini E, Fumagalli E, *et al.* Preoperative imatinib mesylate for unresectable or locally advanced primary gastrointestinal stromal tumors (GIST). *Eur J Surg Oncol* 2009;**35**:739–45. doi:S0748-7983(08)01845-3 [pii] 10.1016/j.ejso.2008.11.005

10 Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA* 2003;**290**:1624–32. doi:10.1001/jama.290.12.1624

11 van Staa TP, Goldacre B, Gulliford M, *et al.* Pragmatic randomised trials using routine electronic health records: putting them to the test. *Br Med J* 2012;**344**. doi:Artn E55Doi 10.1136/Bmj.E55

12 CDISC. ESDI (eSDI) G-. Leveraging the CDISC Standards to Facilitate the use of Electronic Source Data within Clinical Trials. 2006.

13 Kuchinke W, Karakoyun T, Ohmann C, *et al.* Extension of the primary care research object model (PCROM) as clinical research information model (CRIM) for the ¿learning healthcare system¿. *BMC Med Inform Decis Mak* 2014;**14**:118. doi:10.1186/s12911-014-0118-2

14 Bain L. Healthcare Link and eSource. *CDISC J* 2011.

15 Hersh WR, Cimino J, Payne PRO, *et al.* Recommendations for the use of operational electronic health record data in comparative effectiveness research. *EGEMS*

(Washington, DC) Published Online First: 2013. doi:10.13063/2327-9214.1018

16 Kush R, Alschuler L, Ruggeri R, *et al.* Implementing Single Source: The STARBRITE Proof-of-Concept Study. *J Am Med Informatics Assoc* 2007;**14**:662–73. doi:10.1197/jamia.M2157

17 Mastellos N, Andreasson A, Huckvale K, *et al.* A cluster randomised controlled trial evaluating the effectiveness of eHealth-supported patient recruitment in primary care research: the TRANSFoRm study protocol. *Implement Sci* 2015;**10**. doi:10.1186/s13012-015-0207-3

18 Rector AL, Nowlan WA, Kay S, *et al.* A framework for modelling the electronic medical record. *Methods Inf Med* 1993;**32**:109–19. doi:93020109 [pii]

19 Lim Choi Keung SN, Zhao L, Rossiter J, *et al.* Detailed Clinical Modelling Approach to Data Extration from Heterogeneous Data Sources for Clinical Research. AMIA CRI 2014. 2014.

20 Beale T. Archetypes: Constraint-based Domain Models for Future- proof Information Systems. *OOPSLA 2002 Work Behav Semant* Published Online First: 2001. doi:10.1.1.147.8835

21 HL7 Standards Product Brief - HL7 Version 3 DAM: Biomedical Research Integrated Domain Group (BRIDG). http://www.hl7.org/implement/standards/product_brief.cfm?product_id=71 (accessed 10 Jan2016).

22 Ethier J-F, Curcin V, Barton A, *et al.* Clinical Data Integration Model. Core Interoperability Ontology for Research Using Primary Care Data. *Methods Inf Med* 2014;**53**. doi:10.3414/ME13-02-0024

23 Chen R, Klein G. The openEHR Java reference implementation project. *Stud Heal Technol Inf* 2007;**129**:58–62.

24 Mastellos N, Bliźniuk G, Czopnik D, *et al.* Feasibility and acceptability of TRANSFoRm to improve clinical trial recruitment in primary care. *Fam Pract*

2015;:cmv102 – . doi:10.1093/fampra/cmv102

25 Kuchinke W, Ohmann C, Verheij RA, *et al.* A standardised graphic method for describing data privacy frameworks in primary care research using a flexible zone model. *Int J Med Inform* 2014;**83**:941–57. doi:10.1016/j.ijmedinf.2014.08.009

26 Fernández-Breis JT, Maldonado JA, Marcos M, *et al.* Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts. *J Am Med Inform Assoc* 2013;**20**:e288–96. doi:10.1136/amiajnl-2013-001923

27 Ethier J-F, Curcin V, Barton A, *et al.* Clinical Data Integration Model. *Methods Inf Med* 2014;**54**:16–23. doi:10.3414/ME13-02-0024

28 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;**20**:986–94. doi:10.1136/amiajnl-2012-001312

29 Ngouongo SMN, Löbe M, Stausberg J. The ISO/IEC 11179 norm for metadata registries: does it cover healthcare standards in empirical research? *J Biomed Inform* 2013;**46**:318–27. doi:10.1016/j.jbi.2012.11.008

Discussion

A learning health system requires close coupling of care delivery, research and knowledge translation. This aims at an even tighter link than B2B originally proposed with emphasis on development of knowledge discovery. This evolution is well described in an article by John Westfall et al. published in the Journal of the American Medical Association in 2007. (Westfall *et al.* 2007) While presenting great potential, it also creates new requirements. One of the core requirements to achieve this is transparent, semantically correct data sharing across every step of the way.

Moreover, B2B has been deployed and tested mostly in hospital settings, with only a few projects like ePCRN exploring new avenues in primary care. As such, an international LHS including primary care has not been previously attempted and the TRANSFoRm project offered an interesting opportunity to explore and test approaches for complex LHS data sharing needs. This section will discuss the advantages and limitations of the design approaches selected for the framework. Other projects intersecting this work will also be discussed.

Data requirements in LHS

Given the multitude of organisations involved (care, research and knowledge translation), copying data to a central location for sharing is not possible for political and governance reasons. This is compounded in context of a LHS including primary care (multiple, fragmented organisations) across multiple countries (each with different regulatory frameworks). (Weber 2015) This same context also precludes data federation as an option. Many organisations already have data repositories and will not change or duplicate them. Both XML and relational databases data sets are present and already part of larger systems to cater to other missions. Finding a single data structure that could fit both LHS needs and the other institutions requirements could not be accomplished.

Data Mediation

As a result, data mediation emerges as the approach of choice to support a LHS including primary care. Data sources to be part of the LHS are a mix of public and private organisations, each with a tight budget. They receive demands to participate in multiple projects, each asking time and resources to develop solutions to provide data to a specific platform. As such, it becomes very clear that a core requirement is to minimize resources asked from the data sources to participate in the LHS, but also, often overlooked, to maintain participation over time.

This last requirement informs the analysis of global-as-view versus local-as-view approaches in data mediation. A previously described, DM is structured around a central, conceptual model used to support query expression for data sharing. It also contains local models describing each local data source. Each pair (the central model and a local model) then needs to be aligned in order to allow a query based on the central model to be translated into a locally executable query. In a GAV system, the central model is a view of the sum or union of each local model. As a result, any change in the local sources can bring a change in the central model and possibly affect mappings of other local sources as there are interactions between sources to create the central model. This would go against the goal of minimising resources required of sources. When discussed with potential participants, there was a strong desire on their side to be able to make changes to their local structure without affecting the whole system in order to to retain their independence. By forcing a change in the central model, it can also have a snowball effect on subsystems relying on the central model. In context of a LHS, this is especially important given the target of achieving a tight coupling between each component.

Local-as-view mediation

Given the above, despite potentially lower performance in query translation and increased mapping complexity, the local-as-view approach was chosen. In LAV, a central model is to be built independently of the sources it will integrate. It is based on the users' view of the domain, independently of the availability of data in sources envisioned to be initially part of the system. Each local source then creates a model (data source model – DSM) to describe its structure and terminological use. Afterwards, mappings starting from the central model to the local model are created. Direct one to one mappings are not always possible so operations like conditions or union are allowed in the mapping model. In order to facilitate translations, only binary functions are allowed. They can be chained using intermediate results as input for another function.

This approach has the virtue of isolating each source. Any change affecting one source does not affect the others. If a local modification is applied to part of one source, only the mappings affected need to be updated. If this cannot be done immediately, they can simply be deactivated with the effect that these data elements will not be visible to the system until remapped. Nevertheless, the unaffected mappings can be kept and used by the system in the meantime. So this approach allows very granular isolation of modifications, benefiting both the sources (less pressure during modifications) and the platform (minimisation of data unavailability).

Nonetheless, once mappings are completed, other benefits also emerge. One of the most common questions from users is about knowing where what type of information is. Once the system is in place, by surveying the mapping models for each source, the system can easily determine what elements from the central model are available in each source by extracting the concepts from the central model present in the mapping model. If a data element from a source is not mapped to a central model element, then it is not visible on the platform. Extracting this for each source makes it possible to create a metadata registry for each source. This registry can dynamically be updated by reanalysing the model set at a regular interval.

Model management

After a DM LAV has been selected, it raises another important question: how to pragmatically manage the various models necessary for the system to function and be maintained overtime. While structural models are omnipresent in DM systems, once must not forget the importance of terminological information to enable efficient data sharing. In addition, Rector previously demonstrated that these two models are interdependent and require binding together to derive the full semantics of data.

Semantic interdependence

Structural and semantic models have traditionally been managed separately. Structural models have been mostly managed using project-specific structures with information being expressed using XML or UML. (Stanford and Mikula 2008; Unified Modeling Language (UML)) The models are kept separate from the terminologies and are not accessible using recognized standards through these systems. On the terminological side, open-source, well developed terminology servers like LexEVS and Bioportal have been available and used in multiple projects. LexEVS now complies with the HL7 CTS 2 standard which facilitates loose interactions between components interacting with projects using LexEVS instances.

One common way to handle this situation, especially in the research world, has been to use metadata registries (MDR). Allowing the creation of data elements with definitions and fixed permissible values, most MDR offer a lot of flexibility. This is appreciated by the users creating data elements as they can create a multiplicity of similar, but not quite identical objects that can suit exactly their needs. Each research project is different and may require slightly different data elements.

On the other hand, a LHS rests on a different paradigm. At the core of it, it aims at sharing data and the common subset of data to be shared between research, care and knowledge transfer includes EHR and clinical repositories sources, but also genomic data. This data is obviously used for care, but also, as mentioned previously to preload information in research forms or to contextualise alerts from decision support systems. It so requires a coherent and comprehensive view of the clinical domain shared by all participants in the system. This does not align well with a MDR where each participant can create personalized data elements.

Moreover, a MDR does not allow a unified approach to the interdependence of these models as the resulting platform might contain three different sub-systems: one for structural models, one for terminologies and one to host the MDR. Moreover, when analysed in context of data mediation, a MDR creates additional challenges. Assuming that the MDR could be seen as the central model for data mediation, each element would require a set of mappings to each local model. This would imply important resources to be invested by each local source to be invested in creating mappings for each slightly different data element.

Some approaches used by projects like the Electronic Medical Records and Genomics (eMERGE) network in the United States involve mapping data elements to existing terminological resources like SNOMED or the NCIt to provide some anchor to a standardized pivot. (Kho *et al.* 2011; Pathak *et al.* 2011) Nevertheless, mappings are not restricted to a single resource and different users can create different mappings based on the same information (e.g. diabetes is present in multiple terminologies and some are not part of the UMLS). The resources used for mapping can also contain ambiguity or incoherence between each resource. Since the LHS does not have control over these terminological resources, it cannot ensure a coherent model to support mediation. Finally, while some degree of binding can happen at the central model level with a MDR, binding at the local level is not addressed.

A unified framework

A new approach was therefore needed to allow optimal binding between both model types in context of DM LAV. Firstly, it binding must happen both at the local but also at the central level in the DM system. Therefore, as a first step, we included explicit links between relevant data structures (e.g. fields or xml element content) and terminologies used to code data in these elements. Secondly, upon further comparison, in context of DM, both structural and terminological models require alignment. The requirements to achieve these are quite similar and so a platform supporting one should theoretically provide a strong basis for the other. (Shvaiko and Euzenat 2011)

We consequently studied existing platforms, especially LexEVS. As previously mentioned, LexEVS is well suited as a terminological server to be used within a LHS. Its distributed modes of access (Java, as well as SOAP and RESTful web services) and standard compliance (HL7 CTS 2) facilitate its insertion in the platform. The fact that it can be deployed specifically for, and under the control of, the project also makes it a good candidate.

We then explored if we could expand its use to support structural models within a DM system. LexEVS is an open-source software and supports customs extensions. At the core LexEVS allows multiple terminologies, expressed in multiple formats (e.g. UMLS, MedDRA, OBO) to be all imported in the server and served uniformly. LexEVS comes with loaders out-of-the-box to support the most common format, but custom loaders can be built very easily. It also supports mapping files to express mappings between the terminologies (more generically, between the models) stored in the server.

Based on the metamodel developed for the DSM, we built a LexEVS loader to import each DSM instance. We took a similar approach to allow the import of the mappings between DSMs and the central model. As a result, the central model, local models, terminologies, binding between each type of model and within each category can all be stored, managed and accessed within a single system, LexEVS. And all these activities can be accomplished using methods proposed by the HL7 CTS 2 standard.

The resulting platform can be seen as a unified structural/terminological framework. The similarity between terminological and structural operations mentioned previously also extend to pragmatic characteristics offered by LexEVS. For example, the server natively supports versioning. In case of a change to the central model, a new version can be created and made available without disturbing operations based on the previous version. Mappings can also be versioned and can limit their validity to specific versions of the central model or DSM. Elements can be de-activated without being deleted by altering the "isActive" property. Multi-language facilities are also built-in greatly, easing the support of international project like TRANSFoRm which was deployed with four countries using four different languages. Content is also clearly separated from the hosting platform itself.

One remaining challenge is the absence of tools to facilitate the creation and validation of the DSM and mapping of instances required to include a source in the system. In TRANSFoRm, the models were created manually by the researchers, in collaboration with the data source personnel. While tool creation was not in the project scope, it is clear that these tools would be required to support an independent, larger-scale LHS.

CDIM ontology

While the unified framework described above brings intrinsic benefits, the exposed surface of the system is the central conceptual model itself. Queries to be issued to the system need to express their semantics through it. Naturally, as mentioned previously, in order to share data, participants in the LHS must first share a coherent view of the domain at hand. It draws a clear parallel with ontologies. While multiple specific definitions have been proposed, they are generally understood to be sharable and reusable representations of knowledge for a domain. (Gruber 1995; Burgun 2006) While multiple constructs have been referred to as ontologies, the group of formal applied ontologies to support the biomedical domains has been growing over the last few years.

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 artefacts, ultimately depending on human perception, cultural imprints and social conventions. (Gangemi *et al.*

2002) 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. In the biomedical domain, the OBO Foundry collection of interoperable ontologies is built upon the realist upper ontology BFO. (Grenon *et al.* 2004; Smith *et al.* 2007) The medical domain might be spontaneously seen as 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), which formalizes a diagnosis as an informational content entity about the health status of a patient. (Ceusters and Smith 2015) This is well exemplified by the growing body of ontologies present in the OBO foundry.

We therefore chose to develop CDIM as a realist ontology with the aim of reusing as many existing, well-developed classes as possible. It uses Basic Formal Ontology (BFO) 1.1 as the foundational ontology, based on BFO's central role in the OBO Foundry. (Grenon 2003) Several OBO Foundry ontologies, including OGMS, the Vital Sign Ontology (VSO) and the Information Artifact Ontology (IAO) were directly imported into CDIM. (Goldfain *et al.* 2011; The Information Artifact Ontology (IAO) is an ontology of information entities based on the BFO) CDIM also integrates classes from other ontologies such as the Ontology for Biomedical Investigations (OBI) and the Gene Ontology (GO). (Ashburner *et al.* 2000; Brinkman *et al.* 2010)

Primary care concepts and temporality in CDIM

While existing initiatives were leveraged, two aspects were identified with incomplete coverage. Firstly, while a large spectrum of the clinical domain applies to both primary care and specialized care, the former relies on concepts not necessarily exclusive to it, but certainly more central to the domain than secondary or tertiary care. One example of such concept is the reason for health care encounter. While the reasons to see an endocrinology specialist focusing his practice on diabetes might be self-evident, primary care physicians tend to see patients with undifferentiated presentations (e.g. abdominal pain instead of appendicitis). We therefore needed to create primary care-oriented classes. When adding such classes, we designed them 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 classes diagnosis, symptom and sign in our ontology so that they could be a reason for health care encounter, since all these entities can bear the reason for health care encounter role. This facilitates interoperability operations with other projects using the same core and intermediate level ontologies.

It also facilitates binding with terminologies. Various terminologies contain terms with varied semantic groupings. For example, the International Classification of Primary Care (ICPC-2) contains both symptoms and diagnoses. CDIM can contain binding information to a value set containing only ICPC-2 symptoms (amongst others) for the “symptom” class, as well as another binding to another value set for diagnosis. When querying for “reason for health care encounter role”, the relevant terminological values for ICPC-2 can be construed as the union of value sets bound to the classes that can bear the role.

On the other hand, design choices were made with the help of the users in regard to some concepts around habits (e.g. level of physical activity/sedentary, dietary habits) or behavioural interventions. While definitely present in the primary care domain, they were not given importance by the users in terms of impact on current research or knowledge transfer activities.

Secondly, temporal aspects are rarely, if at all, covered in the existing ontologies that were imported. On the other hand, temporal information is essential to express biomedical queries as illustrated by the TRANSFoRm use cases, but also in other projects like EHR4CR, a European project focusing on the re-use of hospital clinical data for research. (Coorevits *et al.* 2013) To better address these needs, temporal classes were created where necessary. We tried to use equivalent classes when possible. Anonymous classes could be used in this context but equivalent classes provide a unique identifier (URI) to be used as a target for a CDIM-DSM mapping. For example, the class “diagnostic process conclusion instant” (the moment at which a diagnostic process produces a diagnosis) is expressed as:

– “temporal_instant” and (“has temporal occupant” some “diagnostic process conclusion”)

In the end, over 100 classes were created in the CDIM ontology, including more than 25 to address temporal aspects. Several new object properties and axioms were also created. LexEVS already contains a loader supporting the Web Ontology Language (OWL) formatted files. While the original ontology file imports the necessary ontology (e.g. OGMS) or ontology segments through the MIREOT process, they are all merged into a single OWL file prior to import into LexEVS in order to facilitate the loading process. (Courtot *et al.* 2011)

Advantages of using an ontology as a central model for a LHS

Using an ontology as the central model of our data mediation system brings important advantages. It permits (and can enforce) formal definitions for the expressed concepts as classes, with examples and synonyms. One of the roles of the central model is to allow a clear and unambiguous mapping with the DSM. The high clarity achieved in CDIM facilitates mapping creation with the DSM.

Although the TRANSFoRm requirements for data usage were to return data according to a tabular format and given the fact that no participants in the TRANSFoRm LHS expressed the need to use a Semantic Web approach to share data, no formal tools were developed to present data through a SPARQL endpoint for example. Nevertheless, CDIM being an ontology, it could be used to bridge data for sources using a triple store for example. Frameworks like R2RML (relational databases to RDF mapping language) are now available to expose relational data as a virtual SPARQL endpoint. (R2RML: RDB to RDF Mapping Language Schema)

An often overlooked aspect of data sharing platforms is provenance but the domain has been active over the last few years and has seen the emergence of the Open Provenance Model standard to act as a core framework. (Curcin *et al.* 2014) Provenance is essential as a means to achieve traceability and auditability in a digital infrastructure of data, tools, processes and agents. The novelty of the TRANSFoRm provenance framework is that it links the provenance model to the medical domain model, by means of bridging ontologies, thereby enabling verification with respect to established concepts. 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 a 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.

It is important to note that a realist ontology represents the world according to the best available theories. It is therefore by default an organic, dynamic entity which will grow and evolve. The characteristics of LexEVS described above help the process to be structured and non-destructive to the ongoing activities.

One key area that will likely expand over the next years or even months will be the genomic (and eventually proteomic and metabolic) observations. The TRANSFoRm use-case required mostly single-nucleotide polymorphism (SNP) support. At some point, sequence structural variations and mutations, as well as, gene expression data will be relevant to the researchers and clinicians, and such concepts will also need to be included in CDIM. (Masys *et al.* 2012) However 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.

Terminology mappings

While DM LAV is based on a central model and mappings to local sources, the terminological alignment activities cannot exclusively be based on a central, sometimes called pivot, terminology. Given the international nature of the system, various national and international standards need to be supported. The point of view and the approach chosen to build each terminology differs with the result that a direct, one to one mapping cannot always be created. One of the most salient examples is found in diabetes. Previously, diabetes was viewed as insulin-dependent diabetes mellitus or non-insulin dependent diabetes mellitus. The view later changed to diabetes mellitus type 1 and type 2. Yet, no direct equivalence, valid in each context, can be made between both visions. Some type 2 diabetes cases require insulin as a treatment. Some sub-types of diabetes mellitus are also emerging like latent autoimmune diabetes of adults which can be seen as diabetes mellitus type 1.5!

For some queries, researchers using sources coded with both insulin dependence view and type view might elect to equate type 1 with insulin dependence. For some others, it might not be possible. It is also possible to complement terminological mappings with restrictions based on other clinical characteristics to arrive at the desired patient population.

So in our unified framework, one-to-one and one-to-many mappings are stored in the system and used automatically for query expansion. In case of contextual mappings or uncertain alignments like the diabetes example above, terminological mappings can be reviewed by the user and selected at the time of query building.

TRANSFoRm outcome

The project was successfully implemented in 5 countries. Data sources ranging from EHR and clinical repositories to genomic data sets were successfully queried to support each component and complete the three use cases: epidemiological research, randomized control trial and decision support system. Regarding the RCT, it was implemented with five different EHR vendors. CDIM triplets [CDIM | operator | value] were able to provide sufficient expressiveness to support the required criteria and data extraction queries when inserted in query models.

The system was well accepted by the different participants. Its distributed nature facilitated its adoption by the various parties and the chosen approach was deemed acceptable in regard of the intellectual property of the industrial partners (like EHR vendors).

Other projects and future work[2]

Other initiatives are emerging in the learning health systems field. The definition retained for our work requires coupling the three components previously discussed. Nevertheless, other initiatives have been published and described as LHS while not embracing, at least at the moment, the full LHS cycle. Some focus on binding research and care data. For example, PEDSnet is described as a national paediatric learning health system but focuses on knowledge generation with the help of clinical data. Based on available publications, it does not formally address the knowledge transfer part of the cycle for now. (Forrest *et al.* 2014)

Regional initiatives like Path and Portal are also emerging with similar goals, first and foremost building observational cohorts from multiple centres based on care delivery and administrative data. (Amin *et al.* 2014; McGlynn *et al.* 2014) Others are discussion platforms like the American Society of Blood and Marrow Transplantation Clinical Case Forum, an online secure tool to enhance interaction and communication among hematopoietic cell transplantation professionals. (Barba *et al.* 2015)

[2] The reader is referred to the articles presented previously for a discussion of the projects predating TRANSFoRm. This section will focus on more recent initiatives.

The Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS) is another project aiming at building a LHS. (Mandl *et al.* 2014) It requires data to be loaded in what is described as a “side car” which provides tools based on the I2B2 data warehouse and federated queries are based on SHRINE. (Weber *et al.* 2009; Murphy *et al.* 2010) If another data warehouse already exists, its needs to be replaced by the side car or data needs to be duplicated. The initial goal was to identify patient cohorts from care delivery data in major hospital centres. They are now looking into using the Patient Centred Oriented Research Common Data Model (expressed as a relational model) as a pivot to support querying other sources not structured as I2B2, moving toward data mediation. (PCORnet Common Data Model (CDM) - PCORnet) At this point, knowledge transfer is not part of the system yet but they built a functional interface to allow for patient engagement directly on the platform which is a novelty and holds promises to increase social acceptance for this type of system and also to improve recruitment and patient participation.

The EURECA project aims to support the development of cancer research tools by providing a homogeneous framework to access data from EHR systems and clinical trial systems. (EURECA | Home) Its common model is based on HL7 RIM. Data loading through ETL in the common DW is executed from HL7 V3 messages (that must be generated by the participating sites). It then provides query federation through the platform. It explicitly includes clinical and genomic data but uses pivot terminologies (SNOMED CT, LOINC and HUGO Gene Nomenclature Committee). (Alonso-Calvo *et al.* 2015; Ibrahim *et al.* 2015) As opposed to TRANSFoRm, they have chosen to store terminologies in a semantic repository (Sesame server). It offers an interesting avenue in terms semantic web opportunities, but many challenges remain in using such models expressed with description logic with a view to do reasoning. (Bodenreider *et al.* 2007; Ceusters *et al.* 2007; Schulz *et al.* 2007; Cheetham *et al.* 2015) While this project targets a specific domain (oncology) and as such might not encounter all the challenges of a more generic approach, it covers knowledge transfer with an explicit module regarding guidelines usage. This could provide important information for the future development of this aspect in the LHS. The project is ongoing.

While not labelled as such, other initiatives have a similar goal of leveraging hospital data to support clinical research. The EHR4CR project (Electronic health record for clinical research), partly funded by the pharmaceutical industry, deployed its infrastructure in Europe. (De Moor *et al.* 2015) It was designed to support I2B2 data warehouses, as well as those structured following its common data model based on the «A_SupportingClinicalStatementUniversal» model, a component of the Study Design sub-group, proposed by the HL7 Regulated Clinical Research Information Model (RCRIM) Work Group. (Ouagne *et al.* 2012) Hospitals need to have their data warehouse structured according to either format to participate. Federated queries can then be issued to the system based on a developed eligibility criteria query language shown to allow a good expressivity. (Bache *et al.* 2015) EHR4CR uses a pre/post processing approach where fairly generic mini queries are created to extract groups of related data elements (e.g. diagnosis). (Bache *et al.* 2013) For a given eligibility criterion, the data group is first extracted and then further processed within the local EHR4CR node. This requires moving significant amount of data out of the DW each time but the group mentions post-processing allows more flexibility in query complexity as it is not limited by the relational database query language. This has been tested with I2B2 and the local DW schema. TRANSFoRm has been targeting a larger variety of sources (genomic/clinical; primary EMR/national repositories; relation databases/xml…) and in the context of primary care, but it tested a smaller set of eligibility criteria.

The Observational Medical Outcomes Partnership was a public-private partnership that was established to inform the appropriate use of observational healthcare databases for studying the effects of medical products, with a clear focus on drug safety surveillance. (Ogunyemi *et al.* 2013) A large portion of the project addressed challenges inherent to use of observational study in terms of statistics and analytical processes. But in order to achieve this, large cohorts were necessary and collaboration to share data was sought through its membership. One of the major outcomes of the project is the OMOP common data model. (Overhage *et al.* 2012) Presented as a relational model, it is well described and is supported by the community. ETL scripts are available for many EHRs. (Common Data Model | Observational Medical Outcomes Partnership) This work was done in parallel with TRANSFoRm and the development of CDIM. While focused on drug safety and not expressed as an ontology (so it is missing explicit, actionable relationships between concepts other than keys and belonging to a table), it has been used in many projects and validated and could inform part of the development of CDIM in the future in order to align common concepts to facilitate future interoperability.

Other European projects like the European Medical Information Framework project (EMIF) also aims at creating an environment for efficient re-use of health data. (Bastião Silva *et al.* 2015) EMIF is currently at the pilot project phase in a hospital in Spain. (Mayer *et al.* 2015)

In terms of primary care, one of the most developed project is the electronic Primary Care Research Network (ePCRN). It also focuses on research and also requires all data sources to be structured according to the American Society for Testing and Materials Continuity of Care Record (CCR) information model before being integrated on the platform, thereby being more along the lines of federation.

Overall, tendencies from the projects above can be identified. They mostly target relational data sources and rely on data warehousing with federation. No direct use of EHR data is included (other than to feed a DW) and most require a specific DW to be used (mostly I2B2). Finally, they are mostly targeting integration of hospital data into research workflows, and this might explain part of the differences between our approach as implemented in TRANSFoRm and these projects.

Opportunities

In order to help new participants use the platform, tools will need to be created to facilitate development of the various models necessary for DM. Interesting resources could be used to facilitate the data source model creation. An unofficial relational database "crawler" was created to produce skeletons of relation DSM, similarly for XML documents. Much work has also been done the domain of ontology alignment and some could be used to narrow targets sets during mapping creation. Likewise, lexical similarity and proximity calculations within CDIM and the DSM could be used to try to identify related concepts. An approach an approach similar to the one presented here was presented by Mate et al. as a way to organise and guide ETL activities to an I2B2 DW. (Mate *et al.* 2015) While not using a unified platform, they did create various tools to facilitate mappings between source and the central model. Given the similarity of the two approaches at the structural mapping level, further investigation of the existing tools could indicate synergies between both projects. These tools could also be extended to make use of the structural-terminological bindings found in the unified framework to do data consistency and quality verifications in local sources.

One very interesting initiative ongoing at this time is the Fast Healthcare Interoperability Resources (FHIR) initiative hosted under the HL7 organisation and currently in the Draft Standards for Trial Use 2 stage. (Anwar and Doss 2015) It consists in a set of predetermined data models (called "resources") which are modular and can be extended with local extensions and focuses on data exchange. If EHRs are to implement this widely as an export format for patient data, it might become interesting for a LHS to provide pre-defined mappings between FHIR structured patient extracts and CDIM, thereby lowering resources required by EHRs to participate.

Nonetheless, the most promising next step might be to expand and formally test the system using tertiary and secondary care data sources. This is essential to be able to address health care trajectories. While intra-institution processes and outcome are important, some important questions can only be answered by uniting data from various levels in order to get a clear picture for a patient. It is not rare for example that a patient treated in a highly specialized cancer care centre is later on transferred to his regional hospital or even his primary care practice. In this context, to be able to evaluate long term outcomes from therapies dispensed in the cancer centre, longitudinal data from both hospital and primary care needs also to be taken into consideration. Similarly, important knowledge can be discovered on predisposing factors and disease evolution by assembling data from the initial clinical presentation in the primary care office with the hospital data.

This approach is also essential so that physicians can use the system to explore the care they provide to their patients. To be able to visualise how they treat patient in their practice and how these patients fare in the long run, it is necessary to include care received in external institutions.

Nonetheless, to explore the path outlined above, an important issue will need to be addressed. Patient data linkage is currently largely done on a per-project basis. A previous analysis demonstrated that at a regional level, patients having data in multiple institutions can create complexities in data selection and analysis. (Weber 2013) Moreover, ethical rules governing this are not always explicitly communicated to the community and automated systems that could facilitate linkage coupled with learning health systems remain to be developed.

Conclusion

The concept of the learning health system is fairly new but has gained a lot of traction, including in high audience journal outside of the medical informatics community, for example in the New England Journal of Medicine. (Hamburg and Collins 2010) It is also well represented in the top journal of the medical informatics discipline. The LHS can provide important benefits for clinical care, research and knowledge transfer activities but sharing data and knowledge efficiently and correctly is an important challenge.

The LHS presents important requirements in terms of data interoperability and data sharing. Popular approaches like data warehousing, whilst useful as at an institutional level, cannot be used as the core support for data integration in the LHS given the governance and resources necessary to implement it. Data mediation has compelling characteristics in this context. A local-as-view data mediation approach was successfully implemented through the TRANSFoRm project and supported the required data sharing activities outlined by its use cases.

Even though Rector identified the strong interdependence between structural and terminological models when using health care data, most projects handled them separately. This work presented a novel, unifying approach to address this requirement. From local models to the central model to the queries, the framework supports and facilitates binding between structural and terminological information in an explicit way. This represents a significant departure from the previous strategies for addressing interoperability in translational research, and it has been successfully demonstrated within the context of the clinical research studies of the EU TRANSFoRm project.

This is achieved in part by an innovative use of the LexEVS terminology server, expanding its role to host all the models required by the framework. The added benefits include access to structural models through standardized methods as well as versioning and multi-language capabilities.

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. 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 epigenetics become part of patient care.

The framework presented here can serve as a strong foundation to expand knowledge about the necessary systems to support a LHS, from tools to maintain the system, to integration of hospital sources to patient data linkage.

In the end, the unified framework is flexible and should reduce the integration efforts required from the data sources, thereby lowering the cost of entry of this type of research for smaller institutions, and removing the need for larger institutions to invest in additional data warehousing.

Bibliography

Articles and conferences

Abelló A, Samos J, Saltor F.A framework for the classification and description of multidimensional data models. *Database and Expert Systems Applications*. Springer, 2001, 668–77.

Alonso-Calvo R, Perez-Rey D, Paraiso-Medina S *et al.*Enabling semantic interoperability in multi-centric clinical trials on breast cancer. *Comput Methods Programs Biomed* 2015;**118**:322–9.

Amin W, Tsui FR, Borromeo C *et al.*PaTH: towards a learning health system in the Mid-Atlantic region. *J Am Med Inform Assoc JAMIA* 2014;**21**:633–6.

Anwar M, Doss C.Lighting the Mobile Information FHIR. *J AHIMA Am Health Inf Manag Assoc* 2015;**86**:30–4.

Ashburner M, Ball CA, Blake JA *et al.*Gene ontology: tool for the unification of biology The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.

Ashish N, Ambite JL, Muslea M *et al.*Neuroscience Data Integration through Mediation: An (F)BIRN Case Study. *Front Neuroinformatics* 2010;**4**:118.

Bache R, Miles S, Taweel A.An adaptable architecture for patient cohort identification from diverse data sources. *J Am Med Inform Assoc* 2013;**20**:e327–33.

Bache R, Taweel A, Miles S *et al.*An eligibility criteria query language for heterogeneous data warehouses. *Methods Inf Med* 2015;**54**:41–4.

Barba P, Burns LJ, Litzow MR *et al.*Success of an International Learning Healthcare System in Hematopoietic Cell Transplantation: the American Society of Blood and Marrow Transplantation Clinical Case Forum. *Biol Blood Marrow Transplant J Am Soc Blood Marrow Transplant* 2015, DOI: 10.1016/j.bbmt.2015.12.008.

Barker PW, Heisey-Grove MD.EHR adoption among ambulatory care teams. *Am J Manag Care* 2015;**21**:894–9.

Bastião Silva LA, Días C, van der Lei J *et al.*Architecture to Summarize Patient-Level Data Across Borders and Countries. *Stud Health Technol Inform* 2015;**216**:687–90.

Bernstein JA, Friedman C, Jacobson P *et al.*Ensuring public health’s future in a national-scale learning health system. *Am J Prev Med* 2015;**48**:480–7.

Bodenreider O.The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267–70.

Bodenreider O, Smith B, Kumar A *et al.*Investigating subsumption in SNOMED CT: an exploration into large description logic-based biomedical terminologies. *Artif Intell Med* 2007;**39**:183–95.

Brinkman RR, Courtot M, Derom D *et al.*Modeling biomedical experimental processes with OBI. *J Biomed Semant* 2010;**1**:S7.

Bruland P, Breil B, Fritz F *et al.*Interoperability in clinical research: from metadata registries to semantically annotated CDISC ODM. *Stud Health Technol Inform* 2012;**180**:564–8.

Burgun A.Desiderata for domain reference ontologies in biomedicine. *J Biomed Inform* 2006;**39**:307–13.

Burgun A, Bodenreider O.Accessing and integrating data and knowledge for biomedical research. *Yearb Med Inform* 2008:91–101.

Calì A, Calvanese D, Giacomo GD *et al.*Accessing Data Integration Systems through Conceptual Schemas. In: S.Kunii H, Jajodia S, Sølvberg A (eds.). *Conceptual Modeling — ER 2001*. Springer Berlin Heidelberg, 2001, 270–84.

Ceusters WM, Spackman KA, Smith B.Would SNOMED CT benefit from Realism-Based Ontology Evolution? *AMIA Annu Symp Proc* 2007;**2007**:105–9.

Ceusters W, Smith B.Biomarkers in the ontology for general medical science. *Stud Health Technol Inform* 2015;**210**:155–9.

Cheetham E, Gao Y, Goldberg B *et al.*Formal representation of disorder associations in SNOMED CT. Lisbon, Portugal: Francisco M. Couto and Janna Hastings, 2015.

Cimino JJ.High-quality, standard, controlled healthcare terminologies come of age. *Methods Inf Med* 2011;**50**:101–4.

Coorevits P, Sundgren M, Klein GO *et al.*Electronic health records: new opportunities for clinical research. *J Intern Med* 2013;**274**:547–60.

Corrigan D.An ontology driven clinical evidence service providing diagnostic decision support in family practice. *AMIA Jt Summits Transl Sci Proc AMIA Summit Transl Sci* 2015;**2015**:440–4.

Courtot M, Gibson F, Lister AL *et al.*MIREOT: The minimum information to reference an external ontology term. *Appl Ontol* 2011;**6**:23–33.

Curcin V, Miles S, Danger R *et al.*Implementing interoperable provenance in biomedical research. *Future Gener Comput Syst* 2014;**34**:1–16.

Delaney B.TRANSFoRm: Translational Medicine and Patient Safety in Europe. In: Grossman C, Powers B, McGinnis JM (eds.). *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, 198–202.

Delaney BC, Curcin V, Andreasson A *et al.*Translational Medicine and Patient Safety in Europe: TRANSFoRm-Architecture for the Learning Health System in Europe. *BioMed Res Int* 2015;**2015**:961526.

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**:54–9.

De Moor G, Sundgren M, Kalra D *et al.*Using electronic health records for clinical research: The case of the EHR4CR project. *J Biomed Inform* 2015;**53**:162–73.

Ethier J-F, Curcin V, Barton A *et al.*Clinical data integration model Core interoperability ontology for research using primary care data. *Methods Inf Med* 2015;**54**:16–23.

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 JAMIA* 2013;**20**:986–94.

Forrest CB, Margolis PA, Bailey LC *et al.*PEDSnet: a National Pediatric Learning Health System. *J Am Med Inform Assoc* 2014;**21**:602–6.

Friedman C, Rubin J, Brown J *et al.*Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. *J Am Med Inform Assoc* 2015;**22**:43–50.

Galvin M, Madden C, Maguire S *et al.*Patient journey to a specialist amyotrophic lateral sclerosis multidisciplinary clinic: an exploratory study. *BMC Health Serv Res* 2015;**15**:571.

Gangemi A, Guarino N, Masolo C *et al.*Sweetening Ontologies with DOLCE. In: Gómez-Pérez A, Benjamins VR (eds.). *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*. Springer Berlin Heidelberg, 2002, 166–81.

Goldfain A, Smith B, Arabandi S *et al.*Vital Sign Ontology. *Proceedings of the Workshop on Bio-Ontologies*. Vienna, 2011, 71–4.

Grenon pierre.*Bfo in a Nutshell : A Bi-Categorical Axiomatization of Bfo and Comparison with Dolce*. University of Leipzig, 2003:33.

Grenon P, Smith B, Goldberg L.Biodynamic ontology: applying BFO in the biomedical domain. *Stud Health Technol Inform* 2004;**102**:20–38.

Gruber TR.Toward principles for the design of ontologies used for knowledge sharing? *Int J Hum-Comput Stud* 1995;**43**:907–28.

Hamburg MA, Collins FS.The path to personalized medicine. *N Engl J Med* 2010;**363**:301–4.

Hernandez T, Kambhampati S.Integration of Biological Sources: Current Systems and Challenges Ahead. *SIGMOD Rec* 2004;**33**:51–60.

Huser V, Cimino JJ.Desiderata for Healthcare Integrated Data Repositories Based on Architectural Comparison of Three Public Repositories. *AMIA Annu Symp Proc* 2013;**2013**:648–56.

Ibrahim A, Bucur A, Perez-Rey D *et al.*Case Study for Integration of an Oncology Clinical Site in a Semantic Interoperability Solution based on HL7 v3 and SNOMED-CT: Data

Transformation Needs. *AMIA Jt Summits Transl Sci Proc AMIA Summit Transl Sci* 2015;**2015**:71.

Khnaisser C, Lavoie L, Diab H *et al.*Data Warehouse Design Methods Review: Trends, Challenges and Future Directions for the Healthcare Domain. In: Morzy T, Valduriez P, Bellatreche L (eds.). *New Trends in Databases and Information Systems*. Springer International Publishing, 2015, 76–87.

Kho AN, Pacheco JA, Peissig PL *et al.*Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium. *Sci Transl Med* 2011;**3**:79re1–79re1.

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.

Landman GWD, Kleefstra N, van Hateren KJJ *et al.*Metformin associated with lower cancer mortality in type 2 diabetes: ZODIAC-16. *Diabetes Care* 2010;**33**:322–6.

Lenzerini M.Data integration: a theoretical perspective. *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. New York, NY: ACM, 2002, 233–46.

Lim Choi Keung SN, Ethier J-F, Zhao L *et al.*The integration challenges in bridging patient care and clinical research in a learning healthcare system. San Francisco, 2014.

Louie B, Mork P, Martin-Sanchez F *et al.*Data integration and genomic medicine. *J Biomed Inform* 2007;**40**:5–16.

Mandl KD, Kohane IS, McFadden D *et al.*Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS): Architecture. *J Am Med Inform Assoc* 2014;**21**:615–20.

Mansmann S, Neumuth T, Burgert O *et al.*Conceptual Data Warehouse Design Methodology for Business Process Intelligence. *Complex Data Warehous Knowl Discov Adv Retr Dev Innov Methods Appl Innov Methods Appl* 2009:129.

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**:734–8.

Mastellos N, Andreasson A, Huckvale K *et al.*A cluster randomised controlled trial evaluating the effectiveness of eHealth-supported patient recruitment in primary care research: the TRANSFoRm study protocol. *Implement Sci* 2015;**10**:15.

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**:419–22.

Mate S, Köpcke F, Toddenroth D *et al.*Ontology-based data integration between clinical and research systems. *PloS One* 2015;**10**:e0116656.

Mayer MA, Furlong LI, Torre P *et al.*Reuse of EHRs to Support Clinical Research in a Hospital of Reference. *Stud Health Technol Inform* 2015;**210**:224–6.

McGlynn EA, Lieu TA, Durham ML *et al.*Developing a data infrastructure for a learning health system: the PORTAL network. *J Am Med Inform Assoc JAMIA* 2014;**21**:596–601.

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**:124–30.

Musen MA, Middleton B, Greenes RA.Clinical Decision-Support Systems. In: Shortliffe EH, Cimino JJ (eds.). *Biomedical Informatics*. Springer London, 2014, 643–74.

Nazarenko GI, Kleymenova EB, Payushik SA *et al.*Decision support systems in clinical practice: The case of venous thromboembolism prevention. *Int J Risk Saf Med* 2015;**27 Suppl 1**:S104–5.

Noy NF, Shah NH, Whetzel PL *et al.*BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009;**37**:W170–3.

Ogunyemi OI, Meeker D, Kim H-E *et al.*Identifying appropriate reference data models for comparative effectiveness research (CER) studies based on data from clinical information systems. *Med Care* 2013;**51**:S45–52.

Ouagne D, Hussain S, Sadou E *et al.*The Electronic Healthcare Record for Clinical Research (EHR4CR) information model and terminology. *Stud Health Technol Inform* 2012;**180**:534–8.

Overhage JM, Ryan PB, Reich CG *et al.*Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc JAMIA* 2012;**19**:54–60.

Pathak J, Solbrig HR, Buntrock JD *et al.*LexGrid: A Framework for Representing, Storing, and Querying Biomedical Terminologies from Simple to Sublime. *J Am Med Inform Assoc* 2009;**16**:305–15.

Pathak J, Wang J, Kashyap S *et al.*Mapping Clinical Phenotype Data Elements to Standardized Metadata Repositories and Controlled Terminologies: The eMERGE Network Experience. *J Am Med Inform Assoc* 2011;**18**:376–86.

Pecoraro F, Luzi D, Ricci FL.Designing ETL Tools to Feed a Data Warehouse Based on Electronic Healthcare Record Infrastructure. *Stud Health Technol Inform* 2015;**210**:929–33.

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**:93–7.

Phillips LS, Branch J William T., Cook CB *et al.*Clinical Inertia. *Ann Intern Med* 2001;**135**:825–34.

Qamar R, Kola JS, Rector AL.Unambiguous data modeling to ensure higher accuracy term binding to clinical terminologies. *AMIA Annu Symp Proc* 2007;**2007**:608–13.

Rector AL.Clinical terminology: why is it so hard? *Methods Inf Med* 1999;**38**:239–52.

Rector AL, Qamar R, Marley T.Binding ontologies and coding systems to electronic health records and messages. *Appl Ontol* 2009;**4**:51–69.

Reynolds T.Clinical trials: can technology solve the problem of low recruitment? *BMJ* 2011;**342**:d3662.

Rodger IW.From Bench to Bedside. *Am J Respir Crit Care Med* 2000;**161**:S7–10.

Sarkar IN.Biomedical informatics and translational medicine. *J Transl Med* 2010;**8**:22.

Sarkar IN, Butte AJ, Lussier YA *et al.*Translational bioinformatics: linking knowledge across biological and clinical realms. *J Am Med Inform Assoc JAMIA* 2011;**18**:354–7.

Schulz S, Suntisrivaraporn B, Baader F.SNOMED CT's problem list: ontologists' and logicians' therapy suggestions. *Stud Health Technol Inform* 2007;**129**:802–6.

Shvaiko P, Euzenat J.A Survey of Schema-Based Matching Approaches. In: Spaccapietra S (ed.). *Journal on Data Semantics IV*. Vol 3730. Springer Berlin / Heidelberg, 2005, 146–71.

Shvaiko P, Euzenat J.Ontology Matching: State of the Art and Future Challenges. *IEEE Trans Knowl Data Eng* 2011;**PP**:1.

Smith B, Ashburner M, Rosse C *et al.*The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;**25**:1251–5.

Soler JK, Okkes I, Oskam S *et al.*An international comparative family medicine study of the Transition Project data from the Netherlands, Malta and Serbia Is family medicine an international discipline? Comparing incidence and prevalence rates of reasons for encounter and diagnostic titles of episodes of care across populations. *Fam Pract* 2012a;**29**:283–98.

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* 2012b;**20**:13–23.

Stanford J, Mikula R.A model for online collaborative cancer research: report of the NCI caBIG project. *Int J Healthc Technol Manag* 2008;**9**:231–46.

Taboada M, Lalin R, Martinez D.An Automated Approach to Mapping External Terminologies to the UMLS. *IEEE Trans Biomed Eng* 2009;**56**:605–18.

Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington, D.C.: National Academies Press, 2011.

Vassiliadis P, Simitsis A, Skiadopoulos S.Conceptual Modeling for ETL Processes. *Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP*. New York, NY, USA: ACM, 2002, 14–21.

Weber GM.Federated queries of clinical data repositories: the sum of the parts does not equal the whole. *J Am Med Inform Assoc JAMIA* 2013;**20**:e155–61.

Weber GM.Federated queries of clinical data repositories: Scaling to a national network. *J Biomed Inform* 2015;**55**:231–6.

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**:624–30.

Westfall J, Mold J, Fagnan L.PRactice-based research—“blue highways” on the nih roadmap. *JAMA* 2007;**297**:403–6.

Wiederhold G.Mediators in the architecture of future information systems. *Comput J* 1992;**25**:38–49.

Web Resources

BioPortal SPARQL Query Browser at <http://sparql.bioontology.org/>

CDISC | Strength Through Collaboration at <http://www.cdisc.org/>

Common Data Model | Observational Medical Outcomes Partnership at <http://omop.org/CDM>

CPRD | *Clinical Practice Research Datalink* at <http://www.cprd.com/>

CTS2 - HL7Wiki at <http://wiki.hl7.org/index.php?title=CTS2>

EURECA | Home at <http://eurecaproject.eu/>

ICD-10 Version:2008 - E10 at <http://apps.who.int/classifications/icd10/browse/2008/en#/E10>

ICD-10 Version:2008 - F32 at <http://apps.who.int/classifications/icd10/browse/2008/en#/F32>

ICD-10 Version:2008 - J18 at <http://apps.who.int/classifications/icd10/browse/2008/en#/J18>

LexEVS | *LexEVS - NCI* at <https://wiki.nci.nih.gov/display/LexEVS/LexEVS>

Ontario Cancer Registry | *Cancer Care Ontario* at <https://www.cancercare.on.ca/ocs/csurv/stats/ocr/>

PCORnet Common Data Model (CDM) - PCORnet at <http://www.pcornet.org/pcornet-common-data-model/>

R2RML: RDB to RDF Mapping Language Schema at <http://www.w3.org/ns/r2rml>

Read Codes | *Read Codes - Health & Social Care Information Centre* at <http://systems.hscic.gov.uk/data/uktc/readcodes>

SNOMED CT | *The Global Language of Healthcare* at <http://www.ihtsdo.org/snomed-ct>

The Information Artifact Ontology (IAO) is an ontology of information entities based on the BFO at <http://code.google.com/p/information-artifact-ontology/>

Unified Modeling Language (UML) at <http://www.uml.org/>

WHO | International Classification of Diseases (ICD) | *WHO* at <http://www.who.int/classifications/icd/en/>

WONCA | International Classification of Primary Care | *WONCA* at <http://www.globalfamilydoctor.com/groups/WorkingParties/wicc.aspx>

Table of Figures

Annexe : Synthèse des travaux et résultats en français[1]

La recherche translationnelle

La recherche translationnelle a été précédemment décrite en utilisant l'acronyme anglophone B2B : Bench to the bedside, de la paillasse au lit du patient. Dans ce paradigme, les idées émergent des activités de recherche fondamentale et la recherche translationnelle B2B visait à augmenter l'utilisation des nouvelles approches découvertes par les scientifiques pour les soins aux patients. Le concept a, par la suite, évolué vers la fin des années deux mille pour inclure « back to the bench », où une vision bidirectionnelle était développée. Cette évolution tire en partie son origine dans l'augmentation de la disponibilité des données génotypiques et phénotypiques.

Aujourd'hui, une quantité énorme de données est créée par les équipes traitantes lors de la provision des soins aux patients. Ces données permettent une utilisation pour la fouille de données et la génération de nouvelles hypothèses, au lieu de cantonner cette dernière aux activités de recherche fondamentale. Néanmoins, les soins de santé sont prodigués par un écosystème complexe où les patients interagissent avec plusieurs cliniciens faisant partie de plusieurs institutions pour des épisodes de soins ponctuels (ex. une pneumonie) ou lors de plusieurs visites reliées à un même problème (ex. le suivi de l'hypertension ou du diabète).

Bien qu'initialement surtout disponibles dans les centres hospitaliers universitaires, les données cliniques sont maintenant créées lors de la majorité des actes de soins, et ce, même lors des activités de la première ligne (médecine de ville) dans plusieurs pays. Les patients participent aussi à plusieurs protocoles de recherche où diverses données biomédicales sont créées et stockées sous forme digitale. Finalement plusieurs systèmes experts comme les systèmes d'aide à la décision (SAD) génèrent aussi beaucoup de données lors de leur utilisation. Avec tous ces pôles de création de données, il devient difficile d'obtenir une image claire, complète et unifiée des interactions d'un patient avec le système de santé.

Cette nécessité de lier les données de patients provenant de plusieurs sources hétérogènes mandate le développement de méthodes standardisées afin de représenter

[1] Le lecteur est invité à se référer à la version complète (en anglais) de la thèse pour les références scientifiques pertinentes.

l’information à partager. Les terminologies et les vocabulaires contrôlés jouent un grand rôle à cet effet. Toutefois, plusieurs terminologies disparates peuvent aussi créer de l’hétérogénéité et les initiatives comme l’UMLS (Unified Medical Language System) permettent de relier les terminologies entre elles au niveau conceptuel (équivalence et subsumption). Afin d’opérationnaliser l’utilisation de ces ressources, des outils comme LexEVS et Bioportal ont été développés. Néanmoins, plusieurs défis persistent dans ce domaine.

Différentes caractéristiques du système de santé expliquent cet état de fait. Plusieurs systèmes informatiques coexistent dans une même institution, mais remplissent différents rôles (pharmacie, laboratoire, finances). Plusieurs compagnies offrent des systèmes pour ces activités, mais les données créées sont structurées et modélisées différemment. La granularité varie aussi beaucoup entre les données des médecins de ville, des hôpitaux et des registres populationnels (comme les registres de cancer). En ajoutant les barrières administratives et géographiques (les patients sont parfois traités dans plusieurs pays en Europe), la fragmentation des données devient importante.

Le système de santé inclut aussi les activités de recherche qui s’y déroulent. Les défis de recrutement sont importants et les coûts, substantiels. C’est d’autant plus vrai en contexte de recherche incluant la première ligne où chaque site, un cabinet de médecin, comprend relativement peu de patients. Il est donc difficile, mais essentiel de bien choisir les sites.

Finalement, les activités de transfert des connaissances prennent de plus en plus d’importance. On les retrouve tant au niveau populationnel (ex. politiques de santé publique) qu’au niveau patient (ex. les outils d’aide à la décision). Néanmoins, plusieurs difficultés apparaissent lorsque ces activités de transfert de connaissances ne sont pas bien arrimées avec les autres activités du système de santé et des populations ciblées. Par exemple, plusieurs systèmes d’aide à la décision sont ignorés par les cliniciens étant donné la présence d’alertes non pertinentes trop fréquentes, le phénomène étant appelé « alert fatigue » dans la littérature anglo-saxonne. Afin d’améliorer la pertinence du système, il est essentiel d’obtenir des informations au regard de la population traitée et du patient index à qui s’applique une alerte. Tel que mentionné ci-haut, afin d’avoir une image complète du patient, on ne peut pas se limiter aux données disponibles dans l’institution de référence.

En tenant compte des différentes caractéristiques et des défis rencontrés lors de la distribution des soins de santé, de la recherche et des activités de transfert des connaissances, il apparaît clair que la recherche translationnelle, qui touche à tous ces domaines, doit travailler à développer des outils qui permettent d'unifier l'échange de données dans le système entre ces activités qui sont, au finale, très interdépendantes.

Les systèmes de santé apprenants

Une réponse possible à ces défis a commencé à émerger il y a quelques années sous la forme des systèmes de santé apprenants. Initialement proposé graduellement par McGinnis et Friedman, le concept propose un arrimage serré et cohérent entre la provision des soins, la recherche et les activités de transfert des connaissances.

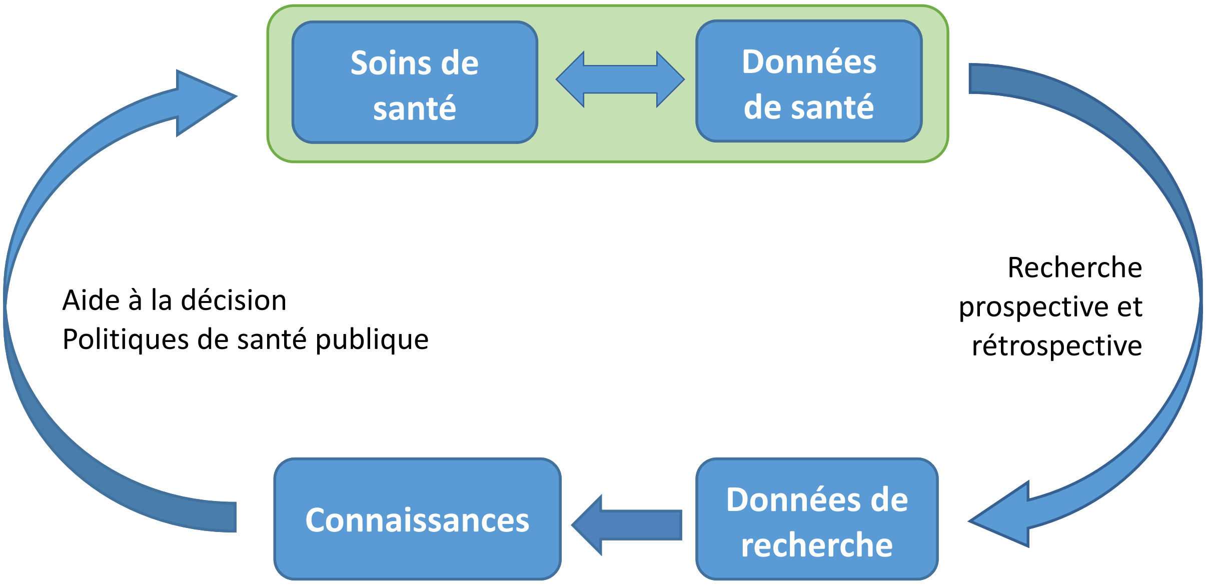


Figure : Système de santé apprenant

Afin de permettre à un cycle similaire de bien fonctionner, une connaissance partagée de la sémantique des données échangées est essentielle et c'est un défi d'interopérabilité important dans un contexte avec autant de sources hétérogènes. De plus, il est maintenant établi que l'étude des maladies complexes requiert l'intégration efficace des données de génomique, de protéomique, ainsi que des données environnementales et phénotypiques.

Le projet Translational Medicine and Patient Safety in Europe (TRANSFoRm), un projet financé par la Commission Européenne, visait à explorer les exigences d'un système de santé apprenant déployé en Europe et supportant les soins de première ligne. Nous avons donc développé notre approche unificatrice et l'avons testée dans le cadre de TRANSFoRm.

Les défis d'interopérabilité biomédicale dans un contexte de systèmes de soins de santé apprenants

Hétérogénéité des sources de données

L'hétérogénéité découle des différentes caractéristiques des sources considérées. Premièrement, la structure de la source de données elle-même peut s'appuyer sur différentes technologies (ex. les bases de données relationnelles, les fichiers Extensible Markup Language – xml – les documents avec valeurs séparées par des virgules – csv…) qui varient en plus selon l'implémentation spécifique préconisée par la compagnie proposant la technologie. De plus, le monde biomédical peut être modélisé de plusieurs façons, selon plusieurs angles, en fonction de l'auteur et des prémisses de modélisation. Finalement, le niveau de granularité peut aussi varier. Donc, même si plusieurs éléments sont largement repris dans le monde de la santé, par exemple la date de naissance ou le numéro de dossier de santé du patient, ils seront souvent modélisés différemment. L'ensemble de ces aspects est identifié dans ce travail comme « hétérogénéité structurelle ».

Deuxièmement, afin de faciliter les échanges d'informations, plusieurs terminologies et vocabulaires contrôlés ont été créés dans le monde biomédical. Ils permettent une utilisation uniforme et constante des termes partagés par les utilisateurs. On peut ici citer des exemples internationaux comme la classification internationale des maladies dixième édition (CIM 10), la classification internationale des soins primaires (ICPC) ou la nomenclature systématisée de la médecine (SNOMED). Néanmoins plusieurs terminologies sont utilisées à un niveau national comme les Read Codes au Royaume-Unis. De plus, les logiciels et plateformes qui sont utilisés dans les systèmes de santé contiennent plusieurs codifications internes (ex. 1 pour les hommes et 2 pour les femmes) qui ne sont pas standard et qui peuvent même être la propriété exclusive des sociétés ayant créé ces logiciels. Globalement, ces aspects sont identifiés ici comme l'hétérogénéité terminologique.[2]

[2] Certains auteurs utilisent le terme « hétérogénéité sémantique » à cet effet, mais tel que démontré dans ce travail, déterminer la sémantique (signification) d'une donnée nécessite l'intégration des aspects structurelles et terminologiques.

Interdépendance des modèles structurels et terminologiques

Les deux types d'hétérogénéité et les modèles sous-jacents peuvent être décrits séparément tels que présentés ci-haut, mais sont au final très interdépendants. Afin de pouvoir déterminer la sémantique (signification) complète d'une donnée, les modèles structurels et terminologiques doivent être intégrés.

Par exemple, le code E11 de la terminologie CIM 10 réfère à l'entrée E11 qui contient plusieurs informations pertinentes pour comprendre la signification du code, incluant son intitulé : Diabète sucré non insulino-dépendant. Les pathologies incluses et exclues sont aussi mentionnées (ex : exclusion du diabète de grossesse). Néanmoins, ceci nous donne seulement une connaissance partielle de ce qui est représenté par cette instance de code E11. Est-ce qu'il représente un diagnostic pour le patient index (ce patient souffre de diabète) ou est-ce l'indication qu'un membre de la famille souffre de diabète ? Est-ce un diagnostic fait à l'admission, alors que l'histoire n'est pas encore très claire, ou est-ce un diagnostic final lors du congé d'une hospitalisation ? Est-ce un diagnostic fait par un interne ou un patron ?

Pour les événements ponctuels, par exemple les pneumonies (J18 dans la CIM 10), le code peut représenter une pathologie existante au moment de l'entrée de la donnée ou plutôt faire référence au fait que le patient a souffert d'une pneumonie dans le passé. Les choses sont encore plus complexes pour les maladies épisodiques comme les dépressions majeures (F32 dans la CIM 10). Un deuxième code pour un même patient 18 mois après le premier peut représenter un suivi de l'épisode de soin initié 18 mois plus tôt ou représenter un tout nouvel épisode de dépression. Ces variations sont d'autant plus importantes lorsqu'on essaie d'intégrer les codes de plusieurs institutions. Il est alors bien hasardeux de faire une extraction simple des codes sans prendre le contexte structurel en compte, car c'est lui qui contient les éléments d'information pour compléter la sémantique des codes terminologiques. Bien que ce défi ait été identifié par A. Rector il y a quelques années pour les informations cliniques, l'ajout des données « omic » incluant l'épigénomique ajoute un niveau de complexité.

Autres exigences

Étant donné que plusieurs institutions doivent participer au système de santé apprenant, chacune évoluant potentiellement dans un environnement légal et administratif différent, une simple copie de toutes les données vers un site central n'est pas possible. De plus, plusieurs institutions ont déjà plusieurs systèmes en place, structurés et encodés d'une certaine façon afin de répondre à leur mission première. Au final, les systèmes de santé apprenants ne contrôlent pas ces organisations et ne peuvent donc pas imposer des modifications à la structure des données existantes ou futures. D'ailleurs, étant donné la quantité importante de projets qui demandent la coopération des organisations de santé avec des budgets limités, il est important de minimiser les ressources nécessaires pour participer à un système de santé apprenant.

Finalement, le système de santé apprenant se doit de supporter les approches prospectives. Contrairement à d'autres domaines, plusieurs requêtes qui devront être exécutées par le système de santé apprenant ne seront pas connues à l'avance. De nouveaux concepts et de nouvelles façons d'interagir avec les données émergeront dans les prochaines années et le système devra être capable de les supporter. De la même façon, toutes les sources ne seront pas connues au jour un. Le système doit donc être capable de croître organiquement et d'intégrer de nouvelles sources en limitant au maximum les impacts négatifs sur le système déjà en place. Conséquemment, l'approche choisie pour gérer l'interopérabilité ne pourra pas être développée statiquement, sur la base des exigences obtenues de focus groupe ou du contenu disponible dans les sources connues au départ.

Revue des approches disponibles en regard de la gestion de l'interopérabilité dans le cadre des systèmes de santé apprenant supportants les soins de première ligne

Les approches pour l'intégration de données peuvent être classifiées en deux grandes familles : les entrepôts de données et la médiation de données. Certaines approches sont aussi présentées comme étant de la fédération de données, ce qui peut être vu comme un cas spécifique de médiation de données où chaque source présente la même structure et utilise les mêmes terminologies pour encoder l'information.

Les entrepôts de données représentent l'approche la plus connue avec des initiatives largement utilisées comme Informatics for Integrating Biology & the Bedside (I2B2) dans le domaine biomédical. C'est une approche souvent utilisée pour intégrer des sources de données sous le contrôle d'une même institution en utilisant le paradigme ETL (Extract Transform and Load) pour charger les données des sources vers un dépôt central. Récemment, les modélisations les plus utilisées étaient les approches multidimensionnelles (ex. flocons ou étoiles), mais d'autres approches sont aussi présentées dans la littérature. Pour plus de détails, le lecteur est invité à se référer à une revue de la littérature récente par Khnaisser et al (2015). Néanmoins, pour un système de santé apprenant cette approche, qui nécessite que les données quittent les barrières institutionnelles pour aller dans un dépôt unique, n'est pas possible. C'est d'autant plus vrai pour le projet TRANSFoRm en contexte européen avec plusieurs pays participants et des législations différentes.

La fédération de données a été utilisée avec succès par plusieurs réseaux de recherche comme BIRN dans le domaine de la neurologie ou le projet ePCRN (electronic Primary Care Research Network). Il s'agit de réseaux distribués, constitués de sources structurées identiquement. Par exemple, dans le cas d'ePCRN, les données doivent être structurées selon le modèle de l'American Society for Testing and Materials continuity of care record. Similairement, basée sur l'initiative I2B2, la plateforme SHRINE permet de fédérer des entrepôts de données I2B2. Étant donné le fait que les sites partagent la même structure, une même requête peut être exécutée à chacun des sites et les résultats peuvent être agrégés facilement par la plateforme centrale. Néanmoins, afin d'obtenir un tel système, les institutions doivent se coordonner et accepter d'utiliser une seule structure. Dans le contexte des systèmes de santé apprenants allant des soins de première ligne aux hôpitaux spécialisés, ce n'est pas possible. De plus, plusieurs institutions ont déjà un entrepôt de données et ne veulent pas le changer ou dupliquer les données, car cela autmenterait leur charge de travail.

Néanmoins, des formes plus génériques de médiation de données sont présentées dans la littérature. La médiation implique avant tout un modèle conceptuel central qui permet d'exprimer les éléments de données nécessaires aux requêtes. Ce dernier est lié aux sources, modélisées elles aussi, par des liens de mise en correspondance (« mappings »). Les requêtes exprimées avec le modèle central sont par la suite envoyées aux sources où elles sont traduites pour être exécutées localement.

Afin de créer les mappings entre le modèle central et les modèles des sources, un niveau doit être construit comme une « vue » de l'autre. On peut donc identifier deux sous-types de médiation de données : « global-as-view » et « local-as-view ». Le premier implique que le modèle central est la vue. Il est donc dérivé étroitement de l'ensemble des sources présentes et de leur contenu. Historiquement, cette approche a présenté de meilleures performances, mais cet aspect ne tient pas compte des avancées technologiques récentes.

Cependant, l’aspect le plus problématique dans le cadre d’un système de santé apprenant dynamique est la dépendance du modèle central par rapport aux sources et donc, une certaine interdépendance des sources à travers les mappings. Si l’une change, le modèle central change et les mappings peuvent potentiellement devoir être ajustés aussi. Dans un contexte où les responsables du modèle central ne contrôlent pas les sources, des incohérences peuvent apparaître rapidement, mais subtilement. Ce n’est donc pas souhaitable dans le cadre d’un système de santé apprenant.

L’approche « local-as-view » dérive plutôt son modèle central des exigences des utilisateurs, indépendamment des données disponibles dans les sources présentes à un temp « t ». Les modèles des sources sont donc plutôt considérés comme des vues du modèle central. Toute donnée des sources ne faisant pas partie des mappings ne sera pas accessible par la plateforme. Néanmoins, certains concepts du modèle central peuvent ne pas être liés à une source sans créer de problème. Le système résultat d’une telle approche représente un modèle stable et pertinent pour les utilisateurs et permet des mappings flexibles et indépendants pour chaque source. C’est donc l’approche qui a été retenue pour développer notre plateforme unifiée d’intégration de données pour les systèmes de santé apprenant.

Même si l’approche de médiation « local-as-view » est une bonne candidate, elle n’a jamais été déployée dans un tel contexte. Bien que des projets comme caBIG aux États-Unis d’Amérique ou Advancing Clinico-Genomic Trials (ACGT) en Europe aient implémenté l’approche dans le domaine de la cancérologie avec des données de recherche en milieux hospitaliers, elle n’a jamais été testée en contexte de soins de première ligne. De plus, les initiatives répertoriées traitent les modèles structurels et terminologiques de façon indépendante. Nous proposons donc ici une approche unifiée pour supporter l’interopérabilité des systèmes de santé apprenants supportant les soins de première ligne basés sur la médiation de données en utilisant une approche « local-as-view ».

TRANSFoRm Project

Le projet TRANSFoRm (http://www.transformproject.eu), qui s’est terminé le 30 novembre 2015, a permis le déploiement d’un système de santé apprenant pour la première ligne en Europe. Trois cas d’utilisation ont été conçus pour orienter son développement.

Le premier consistait à exécuter un protocole de recherche rétrospectif visant à étudier le lien entre certains profils génétiques et la réponse aux sulfonylurées (une classe de médicaments) chez les diabétiques en utilisant les registres cliniques de première ligne. Le deuxième mandatait la mise en place d'une étude randomisée contrôlée sur le reflux gastro-œsophagien et la prise des inhibiteurs de pompe à proton (une classe de médicaments) à partir des données des dossiers électroniques des cabinets de ville dans quatre pays et en utilisant les logiciels de cinq sociétés différentes. Le dernier cas d'utilisation posait son focus sur les systèmes d'aide à la décision pour le diagnostic en maximisant l'utilisation des données contenues dans les dossiers électroniques des cabinets de médecin pour augmenter la pertinence des alertes et leur effet sur l'augmentation de l'acuité diagnostique.

L'article publié dans Biomedical Research International en juin 2015 présente le projet et ses diverses composantes avec plus de détails (Delaney et al. 2015 – voir la thèse principale pour une copie).

Méthode

L'article de Ethier, et al publié dans le Journal of the American Medical Informatics Association en 2013 (voir thèse principale pour une copie), présente la méthodologie générique de l'approche proposée et implémentée dans TRANSFoRm. Pour supporter la médiation, des mappings doivent être créés tant au niveau structurel que terminologique. Pour ce dernier, plusieurs outils standardisés (ex : HL7 Common terminology services 2) existent. Le serveur terminologique LexEVS a été choisi étant donné la possibilité de l'installer localement et donc d'avoir un meilleur contrôle pour le projet. Il présente aussi plusieurs caractéristiques intéressantes comme le support multi-langues (essentiel en Europe) et le support de versions multiples et simultanées, entre autres.

De plus, en analysant les exigences des opérations au niveau des modèles structurels dans le cadre de la médiation de données, des parallèles nombreux sont apparus avec les exigences pour l'utilisation des terminologies. Nous avons donc validé avec succès l'utilisation de LexEVS pour stocker et servir dans un même système les modèles terminologiques et structurels. Cette approche, beaucoup plus efficace et complète, a permis une intégration des deux types de modèles. L'application de contraintes mixtes et inter-reliées entre les deux types de modèles en est aussi facilitée.

Implémentation de l'approche en soins de première ligne

Le modèle central a été implémenté concrètement pour les soins primaires et a été nommé le Clinical Data Integration Model (CDIM). L'article de Ethier et al. publié « ahead of print » en 2014 dans le journal Methods of Information in Medicine décrit l'approche et les principales caractéristiques.

CDIM est une ontologie réaliste basée sur l'ontologie de haut niveau BFO (Basic Formal Ontology). Les ontologies existantes ont été réutilisées au maximum pour favoriser l'interopérabilité (ex. Ontology of General Medical Science, Information Artifact Ontology, etc.). Néanmoins, plusieurs concepts qui sont, sans être exclusifs, essentiels aux soins de première ligne, ne se retrouvaient pas dans les ontologies existantes (ex. « reason for encounter »). Nous avons donc ajouté les classes nécessaires ainsi que deux nouvelles relations. Les aspects temporels sont essentiels afin d'exprimer les requêtes biomédicales, mais ils sont présentement peu exprimés dans les ontologies existantes. Nous avons donc ajouté des classes afin de bien exprimer la temporalité nécessaire pour les requêtes des cas d'utilisation.

Tel que mentionné précédemment, il faut les informations provenant de l'arrimage des modèles structurels et terminologiques pour exprimer une sémantique complète et cohérente. Les concepts dans la plateforme sont donc exprimés sous forme de triplets contenant « identifiant CDIM | opérateur | valeur(s) (ex. code terminologique) ». Bien que certains concepts comme la date de naissance sont naturellement attachés à CDIM et que d'autres très spécifiques comme le pseudohypoparathyroidisme soient plus naturellement orientés vers les terminologies, certains comme les signes vitaux sont dans une zone grise. Les décisions et méthodes de construction sont donc aussi présentées.

Résultats : exemple d'application

L'article récemment soumis à JAMIA en 2016 illustre l'application concrète de la plateforme unifiée. Dans ce cas-ci, l'article présente le fonctionnement de l'intégration de la recherche clinique dans le processus de soins en première ligne. La plateforme permet d'utiliser les standards de l'organisation Clinical Data Interchange Standards Consortium (CDISC), qui sont largement utilisés dans l'industrie de la recherche, pour décrire le protocole et les données nécessaires à sa réalisation. Par la suite, en utilisant CDIM et les données dans le dossier électronique, le système permet l'identification de sujets potentiellement éligibles pour un protocole de recherche en exécutant les critères d'éligibilité. Par la suite, si le sujet est inclus, les formulaires de recherche sont pré-chargés avec les données déjà présentes dans le dossier électronique. Finalement, une copie des formulaires de recherche est envoyée au serveur de recherche alors qu'une autre est stockée dans le dossier électronique.

Le protocole a été déployé avec succès dans quatre pays, cinq différents dossiers électroniques et trois langues différentes.

Discussion[3]

Le concept des systèmes de santé apprenant est relativement nouveau, mais prend de l'ampleur rapidement avec plusieurs initiatives qui émergent. Aux États-Unis d'Amérique, des projets en pédiatrie comme PEDSnet, ou le « Scalable Collaborative Infrastructure for a Learning Healthcare System », qui a une visée plus large du côté adulte, sont présentement en développement. Néanmoins, malgré l'appellation « système de santé apprenant », l'aspect transfert des connaissances est peu ou pas formalisé dans ces systèmes qui visent majoritairement à pouvoir utiliser les données de soins pour la recherche. Du côté européen, le projet Eureca tente de développer une plateforme intégrée pour la cancérologie mais qui inclut nommément le transfert de connaissances, plus spécifiquement des lignes directrices.

D'autres projets comme EHR4CR (Electronic health record for clinical research) ont travaillé dans le domaine de la recherche translationnelle, mais au niveau hospitalier et en supportant des entrepôts de données I2B2 ou du format natif et spécifique au projet. Bien que relié exclusivement au domaine hospitalier, le projet a attentivement étudié plusieurs protocoles de recherche et des centaines de critères d'éligibilité. Le projet TRANSFoRm a, quant à lui, couvert plus de types d'activités (incluant le transfert de connaissances) et de types de sources (xml/relationnel, dossier électronique/registre…), mais beaucoup moins de critères. L'expérience du projet EHR4CR pourrait donc être très intéressante afin d'affiner le modèle de requête et CDIM dans TRANSFoRm.

D'autres projets comme OMOP (Observational Medical Outcomes Partnership) ont été développés dans le cadre spécifique de l'étude de la sécurité des médicaments. Bien que très différent au niveau de l'approche et du domaine, le modèle d'information OMOP, plus

[3] Pour plus de détails, le lecteur est invité à se référer à la section discussion de la thèse principale

restreint que CDIM, a par contre été largement validé et réutilisé par la communauté. Il pourrait donc informer les futurs développements de CDIM.

Au final, des tendances peuvent être identifiées. La majorité des projets visent l'intégration de bases de données relationnelles et s'appuient sur une approche fédérant des entrepôts de données (ex. I2B2 et SHRINE). Il n'y a pas d'utilisation directe des données des dossiers électroniques (autre que pour les transférer dans un entrepôt de données) et plusieurs exigent l'utilisation d'un type d'entrepôt de données précis. Finalement, les projets visent principalement à intégrer les données hospitalières dans le flux de la recherche clinique. Ces particularités peuvent expliquer une partie des différences entre notre approche telle qu'implémentée dans TRANSFoRm et les autres projets.

Conclusion

Les systèmes de santé apprenants prennent de l'ampleur et sont même discutés dans les publications à large lectorat comme le New England Journal of Medicine. Cette vision de la recherche translationnelle mandate des exigences particulières en termes d'interopérabilité. Bien que les entrepôts de données soient souvent utilisés pour l'intégration de données, les systèmes de santé apprenants polyvalents requièrent d'autres approches et la médiation de données a été utilisée avec succès dans le projet TRANSFoRm.

Même si Rector a identifié l'interdépendance forte entre les modèles structurels et terminologiques il y a quelques années, les deux sont encore presque exclusivement traités indépendamment. L'approche présentée par ce travail démontre qu'une plateforme unifiée et basée sur la médiation peut permettre de supporter les activités d'un système de santé apprenant supportant les soins de première ligne, et ce dans plusieurs pays européens.

Ceci résulte en partie de l'utilisation innovante de LexEVS afin d'unifier et de coupler les deux catégories de modèles mais aussi des choix de design pour le modèle conceptuel central, l'ontologie CDIM. La plateforme présentée peut d'ailleurs servir de fondation afin d'explorer plus avant certaines opportunités comme l'intégration plus large de données hospitalières et les outils nécessaires à sa pérennisation.