Detailed Clinical Modelling Approach to Data Extraction from Heterogeneous Data Sources for Clinical Research

Sarah N. Lim Choi Keung, PhD[1], Lei Zhao, MSc[1], James Rossiter, PhD[1], Mark McGilchrist, PhD[2], Frank Culross, MSc, BSc[2], Jean-François Ethier, MD[3], Anita Burgun, MD, PhD[3], Robert A. Verheij, PhD[4], Nasra Khan, MSc[4], Adel Taweel, PhD[5], Vasa Curcin, PhD[6], Brendan C. Delaney, BM BCh, MD[5], Theodoros N. Arvanitis, DPhil[1]
[1]Institute of Digital Healthcare, WMG, University of Warwick, UK;
[2]Health Informatics Centre, University of Dundee, UK;
[3]INSERM UMR_S 872, France;
[4]NIVEL Netherlands Institute for Health Services Research, The Netherlands;
[5]Department of Primary Care and Health Sciences, King's College London, UK;
[6]Department of Computing, Imperial College London, UK

ABSTRACT

The reuse of routinely collected clinical data for clinical research is being explored as part of the drive to reduce duplicate data entry and to start making full use of the big data potential in the healthcare domain. Clinical researchers often need to extract data from patient registries and other patient record datasets for data analysis as part of clinical studies. In the TRANSFoRm project, researchers define their study requirements via a Query Formulation Workbench. We use a standardised approach to data extraction to retrieve relevant information from heterogeneous data sources, using semantic interoperability enabled via detailed clinical modelling. This approach is used for data extraction from data sources for analysis and for pre-population of electronic Case Report Forms from electronic health records in primary care clinical systems.

INTRODUCTION

One of the challenges in healthcare is the efficient reuse of routinely collected data for secondary purposes, such as clinical research. The main uses of electronic health records (eHRs) from patient registries or eHR systems in clinical research are for data analysis and for pre-population of electronic Case Report Forms (eCRFs). While existing patient records can sometimes fulfil all the requirements of a retrospective study analysis, the pre-population of eCRFs from eHRs can cover between 30% and 50% of the requirements[1], and integrated electronic data capture for eCRFs and eHRs can have an even higher overlap, depending on the study[2]. These highlight the potential of reusing clinical data while reducing the amount of redundant data entry (data recorded in clinical care that can be directly used for clinical research). Our research aims to support the interoperability between the clinical researcher tools and the clinical data within patient registries and eHR systems.

The TRANSFoRm project[3] aims to develop rigorous and generic methods for the integration of primary care clinical and research activities, to support patient safety and clinical research. The two clinical research support tools for researchers are the Query Formulation Workbench (QFW) and the eCRF Data Collection Tool. The QFW helps researchers to define studies with eligibility criteria sets for participants, build queries to identify eligible participants, flag patients, and extract data for analysis. The eCRF Data Collection Tool will support primary care practitioners to collect clinical study data and support the collection of patient reported outcome measures (PROMs) via web and mobile methods. In TRANSFoRm, the challenge is to bridge the gap between user requirements in terms of clinical study data items, and the execution of actual queries based on these requirements at the data sources. We adopt a two-level modelling approach[4-6] to separate out the more stable domain information from the various schema implemented by the heterogeneous data sources. The detailed clinical modelling (DCM) approach represents this accurately and will be described further in this paper.

The workflow and the involvement of the TRANSFoRm tools (specifically the QFW) and components are shown in Figure 1, from the definition of the study data extraction requirements to the actual queries at the data sources. In this paper, we focus on cohort identification. Taking the case of a researcher using the QFW to define a retrospective study of patients with Diabetes Mellitus, Step 1 involves defining the data to be extracted from the data sources, without needing to know the format or coding system used in individual data sources. In Steps 2 to 4, a number of TRANSFoRm components are involved to convert the data extract definition into semantically interoperable queries that can be executed at the respective data sources to return the requested data in the format defined by the user.

The remaining sections of this paper are structured as follows to describe DCM approach for semantic interoperability. The Methods section describes the DCM approach as a two-level modelling based on an information model and archetypes to constrain it. The Results section then demonstrates with examples how user requirements are mapped to a specific patient registry schema for data extraction. Finally, we discuss the use of the DCM approach in other TRANSFoRm tools, and finish with some conclusions and future work.

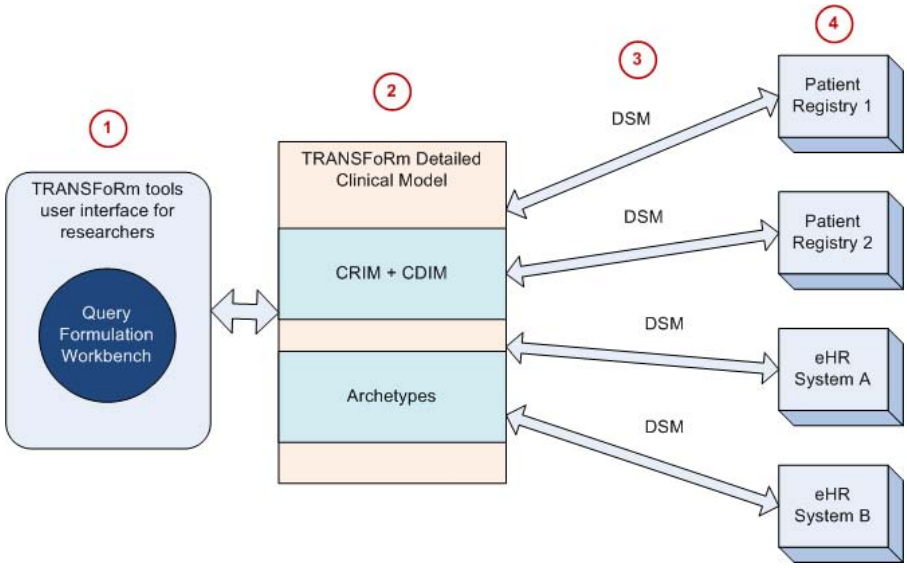

Figure 1. Conceptual workflow, from user definition of data extract requirements to actual queries at data source.

METHODS

Detailed Clinical Models (DCM) organise health information by combining knowledge, data element specification, relationships between elements, and terminology into information models that allow deployment in different technical formats[7,8]. DCM enables semantic interoperability by formalising or standardising clinical data elements which are modelled independently of their technical implementations. The data elements and models can then be applied in various technical contexts, such as eHR, messaging, data warehouses and clinical decision support systems. Work on DCM is still at an early stage with a number of groups involved on an ISO standard for DCM[9].

Within the TRANSFoRm project, the two-level modelling approach of DCM is depicted on the first level as an information model, the Clinical Research Information Model (CRIM), which defines the workflow and data requirements of the clinical research task, combined with the Clinical Data Integration Model (CDIM), an ontology of clinical primary care domain that captures the structural and semantic variability of data representations across data sources. This separation of the information model from the reference ontology has been previously described by Smith and Ceusters[10]. At the second level, archetypes are used to constrain the domain concepts and specify the implementation aspects of the data elements within eHR systems or patient registries. We use the Archetype Definition Language (ADL) to define the constraints and combine them with CDIM concepts in specifying the appropriate data types and range values. The two-level modelling approach, using the concept of archetype for detailed clinical content modelling, has been adopted by ISO/CEN 13606[11,12]. This approach makes it possible to separate specific clinical content from the software implementation. The technical design of the software is driven by the first level information model which specifies the generic information structure of the domain. The archetype defines the data elements that are required by specific application contexts e.g. different clinical studies.

The distributed query and data extraction infrastructure is a central component of the TRANSFoRm software platform. This infrastructure facilitates patient identification and reuse of routine healthcare data for research analysis. The TRANSForm platform interacts with disparate patient registries and eHR systems via the Data Node Connector, which translates the user queries, such as a data extraction definition as part of a retrospective study, in the form of archetypes to data source queries using the Semantic Mediator. The Semantic Mediator ensures the semantic translation queries from the Query Formulation Workbench to individual data source schema with the help

of data source models (DSM) and mappings to CDIM (CDIM-DSM)[13,14]. The transformed query can then be executed at the data source side and results are returned to the user. While specific DSM and CDIM-DSM mappings are required for each data source, these have to be built only once per data source. Additionally, the detailed clinical model is flexible enough to enable researchers to query heterogeneous datasets without any knowledge of the underlying structure, as they themselves do not use the DSM and CDIM-DSM mappings directly.

RESULTS

The data extraction for analysis was carried out for a Diabetes study, using a patient registry sample. In this section, we demonstrate how the data extract definition was processed, from the user at the Query Formulation Workbench, via the TRANSFoRm DCM to the data source. Following the steps in the conceptual workflow in Figure 1, we describe one specific data extract requirement – prescription dates for Metformin medication – for illustration. The clinical researcher defines what data to extract using the Query Formulation Workbench. In the case where the researcher wants to extract all the instances when patients have been prescribed Metformin (Figure 2), the data elements *Medication* and *Prescription date* are selected for extraction, and the constraint on the Medication concept is specified as part of the archetype specification. For example, the researcher can choose Metformin with the ATC code 'A10BA02' from the TRANSFoRm terminology service[15]. The resulting archetype definition in ADL is shown in Figure 3.

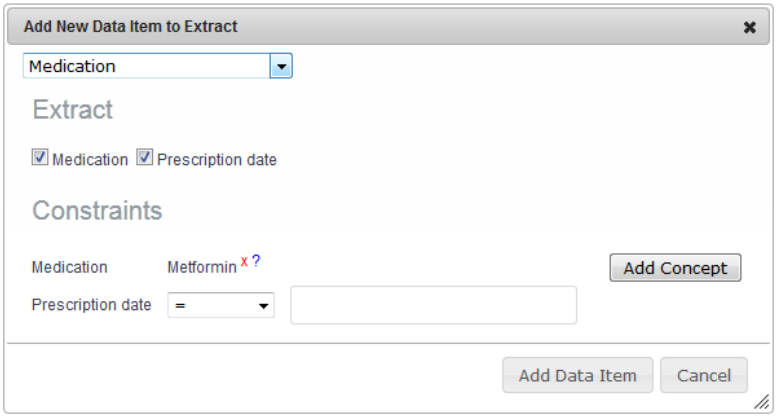

Figure 2: Data extract definition using the Query Formulation Workbench

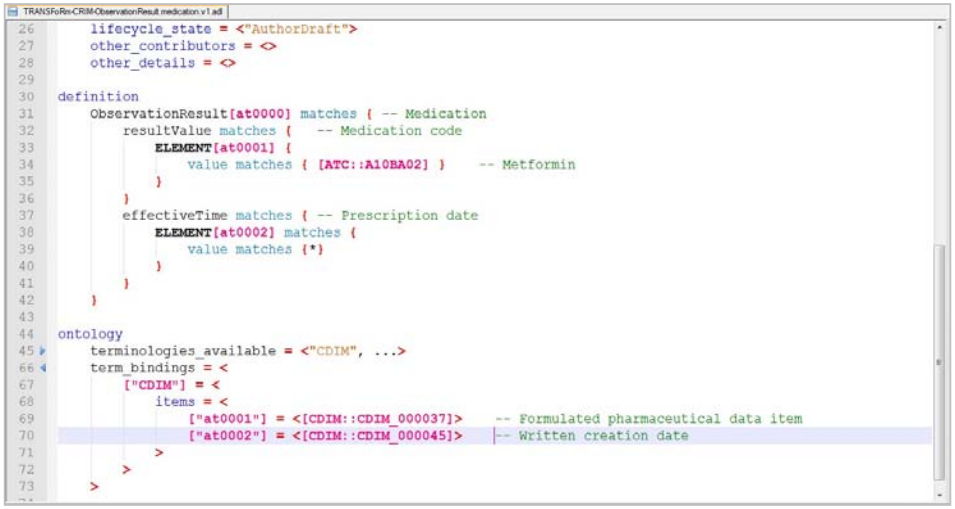

Figure 3: Medication archetype definition in ADL.

The translation of archetypes into a computable form at the data source includes the use of a DSM (Figure 4a) and the CDIM-DSM mappings for the data source (Figure 4b). The DSM defines how the data source organises the medication prescription information, while the CDIM-DSM mappings express information in the form of triplets (CDIM concept; operator; terminology code). For instance, for Metformin with ATC code 'A10BA02', the information triplet is represented as (medication agent; =; 'A10BA02'). Following the transformations, an SQL query is generated to enable the specified data to be extracted from the data source (Figure 5).

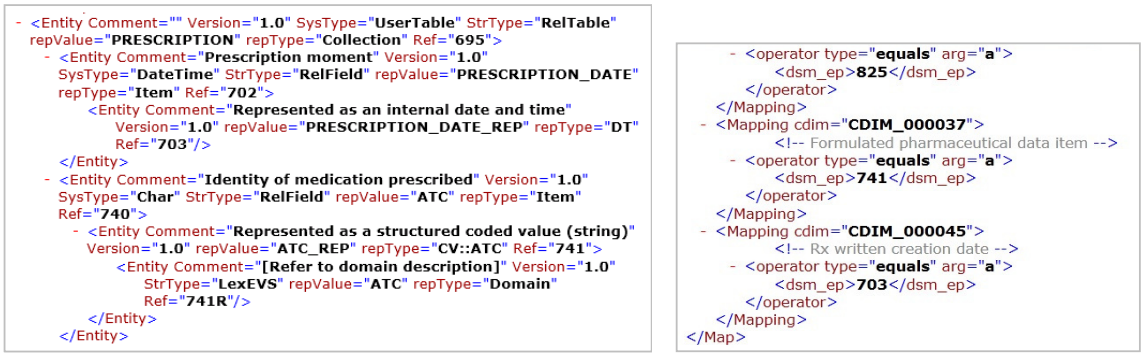

Figure 4: (a) Part of DSM definition **(b)** Part of CDIM-DSM for medication.

```
SELECT PATIENT.ID_PATIENT AS CDIM_000003, PRESCRIPTION.ATC AS CDIM_000037,
PRESCRIPTION.PRESCRIPTION_DATE AS CDIM_000045
FROM PRESCRIPTION
INNER JOIN PATIENT ON PRESCRIPTION.ID_PATIENT = PATIENT.ID_PATIENT
WHERE PRESCRIPTION.ATC IN ('A10BA02')
```

Figure 5: SQL query generated for data source schema.

DISCUSSION

Different solutions have been developed internationally to support a more rapid translation of scientific discoveries into clinical practice, notably i2b2[16]. i2b2 is a data warehousing system that extracts, transforms and loads data into a common schema. In comparison, the TRANSFoRm infrastructure adopts a model-based mediation approach, allowing the querying of heterogeneous data repositories without needing them to be in a single common schema. The TRANSFoRm project also aims to support clinical research with the reuse of eHR data within eCRFs, to avoid duplicate data collection. A minimisation of transcription errors and time-saving are added benefits for the reuse of routinely-collected clinical data. For instance, Köpcke et al.[17] report that the pre-population of case report forms decreased the time for data collection by nine-fold, from a median of 255 to 30 s. The DCM approach can be used in a similar way for the automatic pre-population of eCRFs from eHR systems as for the data extraction for retrospective studies from patient registries. The pre-populated data can be exported in the Operational Data Model (ODM) format[18], a standard for the interchange of data and metadata for clinical research, especially data collected from multiple sources. This will make the pre-populated data compatible with the remaining eCRF and PROM data that are collected as part of a study.

TRANSFoRm uses archetypes in the current implementation as ADL is a user-friendly language and can be easily understood by clinical researchers. HL7 templates, which constrain the HL7 clinical statement pattern, provide an alternative way to implement DCM in the context of HL7[8]. Future improvements to the TRANSFoRm GUI tools can include an authoring tool to assist users in defining new data elements. Referring to the medication archetype definition in Figure 2, currently, a user cannot directly update the archetype structure, for example to add the constraint of the dosage of the medication. Additionally, the tool can support various data element specification formats, such as HL7 templates and archetypes, for interoperation with systems that use these technologies.

CONCLUSION

The reuse of routinely collected data from clinical care in clinical research is an important goal of the TRANSFoRm project. The approach is to retrieve relevant data elements from the data sources (patient registries and eHR systems) without using a common structure to enable interoperability. Researchers can use the TRANSFoRm tools to define their studies without being aware of the underlying structure of the heterogeneous datasets. We have presented how a detailed clinical modelling approach is used to enable semantic interoperability between the researcher-defined queries and the individual data sources. The two-level modelling supports the flexibility of specifying new archetypes, as well as to add new data sources, while keeping the information model stable. Therefore, the DCM approach facilitates the bridging of the gap between clinical research and clinical care. The next steps include the validation of this approach and the related TRANSFoRm tools and components. Validation is being planned based on two use cases, a retrospective genotype-phenotype diabetes study and a prospective study for the gastro-oesophageal reflux disease randomised control trial.

Acknowledgements

The TRANSFoRm project is partially funded by the European Commission under the 7th Framework Programme (Grant Agreement 247787).

References

1. El Fadly A, Rance B, Lucas N, Mead C, Chatellier G, Lastic P-Y, et al. Integrating clinical research with the Healthcare Enterprise: From the RE-USE project to the EHR4CR platform. J Biomed Inform. 2011 Dec;44, Supplement 1:S94–S102.
2. Zahlmann G, Harzendorf N, Shwarz-Boeger U, Paepke S, Schmidt M, Harbeck N, et al. EHR and EDC Integration in Reality [Internet]. Appl. Clin. Trials. 2009 [cited 2013 Oct 1]. Available from: http://www.appliedclinicaltrialsonline.com/appliedclinicaltrials/article/articleDetail.jsp?id=641682
3. TRANSFoRm [Internet]. [cited 2013 Sep 30]. Available from: http://www.transformproject.eu/
4. Rector AL, Nowlan WA, Kay S, Goble CA, Howkins TJ. A framework for modelling the electronic medical record. Methods Inf Med. 1993 Apr;32(2):109–19.
5. Johnson SB. Generic data modeling for clinical repositories. J Am Med Inform Assoc. 1996;3(5):328–39.
6. Beale T. Archetypes: Constraint-based domain models for future-proof information systems. Seattle, Washington, USA, November 4, 2002; 2002. Available from: http://www.openehr.org/files/resources/publications/archetypes/archetypes_beale_oopsla_2002.pdf
7. Goossen W, Goossen-Baremans A, van der Zel M. Detailed Clinical Models: A Review. Health Informatics Res. 2010;16(4):201.
8. Goossen WTF, Goossen-Baremans A. Bridging the HL7 template - 13606 archetype gap with detailed clinical models. Stud Health Technol Inform. 2010;160(Pt 2):932–6.
9. European Committee for Standardization CEN. CEN/TC 251 - Standards under development [Internet]. [cited 2013 Oct 1]. Available from: http://www.cen.eu/CEN/Sectors/TechnicalCommitteesWorkshops/CENTechnicalCommittees/Pages/WP.aspx?param=6232&title=CEN%2FTC+251
10. Smith B, Ceusters W. HL7 RIM: an incoherent standard. Stud Health Technol Inform. 2006;124:133–8.
11. EN 13606 Association. The CEN/ISO EN13606 standard [Internet]. [cited 2013 Oct 2]. Available from: http://www.en13606.org/the-ceniso-en13606-standard
12. Muñoz P, Trigo J, Martínez I, Muñoz A, Escayola J, García J. The ISO/EN 13606 Standard for the Interoperable Exchange of Electronic Health Records. J Healthc Eng. 2011 Mar 1;2(1):1–24.
13. Ethier J-F, Dameron O, Curcin V, McGilchrist MM, Verheij RA, Arvanitis TN, et al. A unified structural/terminological interoperability framework based on LexEVS: application to TRANSFoRm. J Am Med Inform Assoc. 2013 Jan 9;20(5):986–94.
14. Ethier J, McGilchrist M, Burgun A, Sullivan F. D6.3 Data Integration Models [Internet]. 2013. Available from: http://transformproject.eu/TRANSFoRmproject.eu/Deliverables_files/TRANSFoRm%20D6%203%20Data%20Integration%20Models.pdf
15. Lim Choi Keung SN, Zhao L, Tyler E, Arvanitis TN. Integrated Vocabulary Service for Health Data Interoperability. Fourth International Conference on eHealth, Telemedicine and Social Medicine (eTELEMED 2012). Valencia, Spain: IARIA; 2012. p. 124–7.
16. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)[J]. J Am Med Inform Assoc. 2010, 17(2): 124-130.
17. Köpcke F, Kraus S, Scholler A, Nau C, Schüttler J, Prokosch H-U, et al. Secondary use of routinely collected patient data in a clinical trial: an evaluation of the effects on patient recruitment and data acquisition. Int J Med Inf. 2013;82(3):185–92.
18. CDISC. ODM: Operational data Model [Internet]. [cited 2013 Oct 2]. Available from: http://www.cdisc.org/odm