TRANSFoRm: Implementing a Learning Healthcare System in Europe through Embedding Clinical Research into Clinical Practice

Sarah N. Lim Choi Keung, Lei Zhao, Theodoros N. Arvanitis
Institute of Digital Healthcare, WMG, University of Warwick, UK
{s.n.lim-choi-keung, lei.zhao, t.arvanitis}@warwick.ac.uk

Vasa Curcin, Brendan Delaney
Department of Primary Care and Public Health Sciences, King's College London, UK
{vasa.curcin, brendan.delaney}@kcl.ac.uk

Jean-François Ethier, Anita Burgun
INSERM URM 1138, Université Paris Descartes, France
ethierj@gmail.com, anita.burgun@egp.aphp.fr

Mark McGilchrist
Division of Population Health Sciences, University of Dundee, United Kingdom
m.m.mcgilchrist@dundee.ac.uk

Piotr Bródka, Włodzimierz Tuligłowicz
Institute of Informatics, Politechnika Wrocławska, Poland.
{piotr.brodka, wlodzimierz.tuliglowicz}@pwr.edu.pl

Anna Andreasson
Centre for Family Medicine, Karolinska Institutet, Sweden.
anna.andreasson@ki.se

Abstract

Evidence-based practice has become the cornerstone of high quality clinical care. However, producing evidence of sufficient quantity and quality is hampered by the high cost and complexity of conducting research, the numbers of clinicians and patients willing to participate in clinical trials. The Learning Healthcare System initiative creates routes for knowledge transfer between different parts of the health system, thereby increasing its research and learning capacity. This effort is closely aligned with initiatives such as CDISC that are working on standards for computational clinical trial definitions. TRANSFoRm project builds upon existing state-of-the-art technologies for reusing clinical information routinely captured in Electronic Health Record systems, to facilitate conduct of clinical trials in primary care. This paper presents the modeling part of the TRANSFoRm method for building such a Learning Healthcare System and addressing the integration of clinical trials into the routine clinical practice, thereby reducing cost and complexity of the task.

1. Introduction

Randomized controlled trials (RCTs) are the most reliable approach to generating high quality medical evidence. However, protocols for RCTs are slow to produce and often imprecisely specified, resulting in inefficient recruitment and execution [28]. This problem is exacerbated in environments where dedicated research coordinators would have to be widely spread across sites, such as community or practice-based research. Even the pharmaceutical industry, with dedicated staff and software tools, struggles to set up and manage multi-site configurations that are required in order to run studies in the primary care setting [4]. Furthermore, no clinical trial-authoring tool currently captures a unified trace of both the trial design and execution processes, preventing questions directly linking elements of trial design to trial performance, and investigation of best practice in study design.

Resolving this crisis in clinical trials is one of the main objectives of the Learning Healthcare System (LHS), an international initiative that aims to establish a next-generation healthcare system, "... one in which progress in science, informatics, and care culture align to generate new knowledge as an ongoing, natural by-product of the care experience, and seamlessly refine and deliver best practices for continuous improvement in health and health care" [23]. Each participant in the LHS, be they a clinician, patient, or researcher acts as both a consumer and producer of knowledge, with the LHS providing: a) routine and secure aggregation of data from multiple sources, b) conversion of data to knowledge and c) dissemination of that knowledge, in actionable form, to all who can benefit from it [13].

Multiple LHS-related research efforts are currently focused on integration of clinical research and patient care workflows through Electronic Health Records (EHR) systems, which is seen as a way of providing a low-cost method for conducting large-scale pragmatic

1530-1605/15 $31.00 © 2015 IEEE
DOI 10.1109/HICSS.2015.383

clinical trials [2,30]. Important advances have been made in EHR-based phenotyping [17], data exchange between EHR and electronic Case Report Form (eCRF) components [19], trial scheduling [32], and automating the audit trail through research provenance [9]. With 98% of GP practices in the UK using Electronic Health Records and similar figures in other European countries, a number of EU projects are active in the field, such as TRANSFoRm [29], EHR4CR [10] and Semantic Health Net [26]. Several standardization efforts in clinical trial management and reporting are currently in progress, particularly in the US, through Clinical Data Interchange Standards Consortium (CDISC) [5], which sees this integration of clinical and research workflows as a key next step for their standards [18].

While a number of elements have been put in place to secure the vision of EHR-driven clinical trials, a key component that is still missing from the equation are common, usable and easily accessible standards and tools to facilitate design and execution of such trials.

This paper presents a method for achieving semantic interoperability of eCRF data and definitions with routinely collected data residing in EHR systems, to enable pre-population of eCRFs from routine data. This is performed through embedding of TRANSFoRm process and data interoperability models into standards CDISC models that define the study structure.

In the following sections, we will briefly describe the TRANSFoRm project and the GERD study use case in sections 2 and 3. We describe in detail our eCRF modelling methodology in section 4, using the GERD use case as one illustrative example to elaborate the technical details. Next, section 5 details the clinical modeling and the use of archetypes to represent clinical data elements in primary care. Finally, we discuss some related and future considerations in section 5 and conclude in section 6.

2. TRANSFoRm: Implementing LHS in Europe

FP7 TRANSFoRm is an EU funded large scale project, which aims to develop and evaluate a LHS for European Primary Care. The project has three broad aims, to facilitate multiple site genotype-phenotype studies, to enable multi-site, practice-based RCTs by embedding distributed trial functionality into existing EHR systems, and to prototype a diagnostic decision support system linked to EHRs.

A core output of the project is the specification and demonstration of a 'functional' eCRF, designed to enable the collection of semantically controlled and standardized data from within an EHR system. Based on the use case descriptions developed by clinical researchers in TRANSFoRm, the eCRF specification needs to satisfy the following requirements:

1. The eCRF specification will support the collection of semantically–controlled data. Primary care patient data in European EHR systems are generally coded with specific clinical coding schemes such as ICPC, ATC, Read Codes, etc. The eCRF specification should enable the collection of EHR data coded in different coding schemes from various EHR systems in different countries.
2. The development of the eCRF specification will follow a model-based approach in compliance with the project philosophy of TRANSFoRm. An eCRF model will be developed within the framework established by Clinical Research Information Model (CRIM) [20].
3. The eCRF model will enable semantic interoperability between the study system and the EHR system, and support automated pre-population of eCRFs from EHR data.
4. The eCRF model will support the collection of data through a variety of technologies: web interface, mobile application, as well as data collection software embedded in an EHR system.
5. The eCRF specification will support international languages in addition to English. So when deployed in different countries, the case report forms will present native languages to users.
6. The eCRF model will cover the whole lifecycle of study data collection, and include support for various study activities, such as eligible patient identification and recruitment, informed consent, randomization, adverse event monitoring, etc.
7. The specification should have appropriate representations (such as XML) for deployment within an EHR system.
8. The eCRF model will be closely aligned to existing standards produced by CDISC [5] and will be taken to CDISC for incorporation into standards. This is essential to enable interoperability between TRANSFoRm and Clinical Trial Data management and electronic remote data capture systems.

EHR systems across Europe use heterogeneous systems to record and manage clinical data in primary care. Clinical concepts can be semantically interoperable via common ontologies, and mappable medical vocabularies, as has been done in the TRANSFoRm project with the CDIM ontology and the TRANSFoRm vocabulary service. In TRANSFoRm, the clinical data from EHR systems and data warehouses need to be additionally semantically

interoperable with other systems, such as the study system, in particular the eCRF forms and the data that are collected within them, that are common to both the clinical and research domains. eCRFs are commonly structured according to standards, such as CDISC. There is a need to make data elements that are common to both domains semantically interoperable. This is to ensure that data can be extracted from EHR systems to populate eCRFs, and also to report eCRF data back into the EHR systems to assist medical practitioners in primary care if applicable.

In TRANSFoRm, this semantic interoperability between the clinical and research domains is enabled by semantically extending the CDISC ODM standard to maintain the meaning of clinical data elements between systems crossing these two domains.

The eCRF model is currently being validated by the TRANSFoRm Gastroesophageal Reflux Disease (GERD) use case. GERD is a spectrum of disorders mainly caused by the retrograde flow of gastric acid from the stomach into the esophagus. The main symptoms are heartburn and acid regurgitation. GERD is treated using proton pump inhibitors (PPI), H2-blockers or antacids. As many patients require ongoing PPI treatment, continuous versus on-demand PPI use is being compared in this study with outcomes symptom severity and quality of life.

The aim of GERD evaluation study is to: (1) determine the effectiveness in terms of symptom control and quality of life (QoL) and event-initiated assessment of QoL and symptoms in patients with GERD, randomized to either continuous or on-demand use of PPI; and (2) to investigate whether using an electronic system to recruit study participants and to collect data increases the recruitment rate in primary care clinical studies. The study will have the following timeline (Figure 1) and data collection workflow:

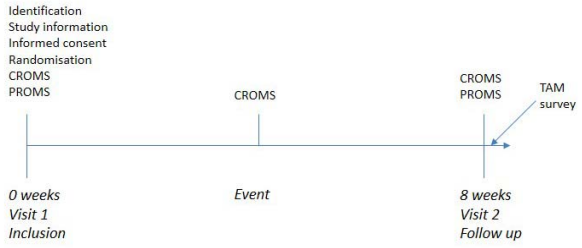

Figure 1. GERD study timeline

Visit 1: The patient attends the primary care practice, either already taking PPI, or as a new diagnosis of GERD. At this point, the patient will be flagged as potentially eligible for the study. The general practitioner (GP) checks the eligibility criteria and gives study information if the patient is eligible. If the patient consents to take part in the study, the patient is randomized to either continuous or on-demand PPI.

- *Randomization:* if the patient consents, the GP confirms the age and gender. This information is used to allocate patients to randomization blocks in the randomization. After completing this point, the patient is randomized (but the outcome is hidden from the GP until the end of the consultation when the dose is prescribed to avoid biasing the GP).
- Care reported outcome measures (CROMs) are completed by the GP in the eCRF. They represent all types of objectively reported care outcomes, collected at the physician's office including data from the patient's electronic health record, such as BMI, blood pressure, and laboratory data.
- GP is informed of the randomization outcome.
- Patient reported outcome measures (PROMs) are completed by the patient via their smartphone (using Android or iOS mobile application) or a web browser. PROMs represent all types of subjective reported outcomes, such as health-related quality of life, treatment satisfaction, subjective health status and subjective symptoms.

Visit 2: After 8 weeks, the patient returns for a follow-up visit. CROMs and PROMs are completed by the GP and patient respectively.

Event: If the patient visits the practice between Visit 1 and Visit 2, CROMs for event visits are completed by the GP.

Study end: participants and GPs will be asked to answer some question on the acceptance of the TRANSFoRm system through a Technology Acceptance Model (TAM) survey [31].

3. Structuring the clinical trial

In order to enable standardized interaction between clinical trial instruments and the EHR systems, the development of the study specification, including eCRF data collection instruments needs to follow a model-based approach. The modelling process in TRANSFoRm is based on CDISC standard information models with the study workflow modelled in the Study Design Model (SDM) [7], and eCRFs modelled by Operational Data Model (ODM) [8]. The CDISC foundational standards form the basis of a suite of standards supporting the clinical research process from protocol through data collection, data management, data analysis and reporting.

Automated data collection from EHR systems requires computable definitions of clinical data elements used. The openEHR archetype approach is used as the basis for a query model to enable formulation of eligibility criteria and data extraction

requests to be executed by the EHR systems. CDISC ODM is extended to embed these query requests, so that the eCRFs can be partially pre-populated from the EHR data.

The file format that will be used to exchange eCRF data between the TRANSFoRm Study System and the EHR systems is the XML representation of SDM, comprising metadata definition, study design and queries and archetypes.

3.1. Study specification model

The clinical research study protocol is the plan that describes the study's objectives, methodology, statistical considerations, and the organization of the study. This plan includes the design of the study, with the arm descriptions, the schedule of activities, the eligibility criteria and summary information. Several CDISC standards have been developed to represent different aspects of the clinical study. The CDISC Study Design Model (SDM) [7] captures the study specification, but not its execution, data collection or reporting. It consists of three groups of elements: Structure (epochs, arms, cells, segments, activities), Workflow (decision points, branches) and Timing (when activities should happen).

We shall now illustrate how this model is used on the example of the GERD study.

Structural Elements define the overall structure of the study. The elements describing parts of the GERD study are illustrated in Figure 3 and described as follows

Epoch: a sequential block of time for a study, e.g. the GERD study has two epochs for treatment and follow-up.

Arm: defines a study arm. The GERD study has two arms for continuous and on-demand PPI.

CellDef: a study cell represents the intersection of an epoch and an arm. For example, the left study cell represents the treatment epoch, which occurs before randomization into continuous or on-demand arms.

SegmentDef: a segment has a set of activities. The continuous follow-up segment depicts the activities involved in the follow-up epoch of the continuous arm for the GERD study.

ActivityDef: an activity represents a point in the study at which a specific action is to be taken. There are 8 activities in the treatment segment depicted in Figure 2, such as trial start, inclusion/exclusion criteria check and randomization.

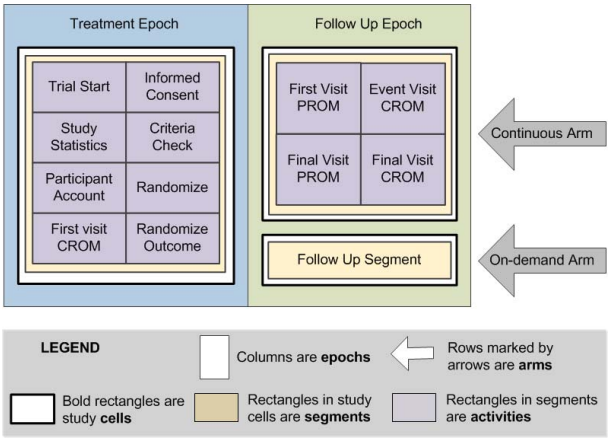

Figure 2. Structural elements of the GERD study

Workflow Elements specify the possible participant paths through a study. They are defined separately from structural elements in the XML file, but they reference objects defined in the Structure section of the document. The main types are: (1) Activities with workflow-specific roles (StudyStart, StudyFinish, PathCanFinish); (2) Entry and exit criteria for activities (EntryCriteria, ExitCriteria, Criterion); (3) Workflow paths and branching (Transition, Switch, Trigger). Figure 3 shows Transition and Switch elements for eligibility criteria check activity.

```
<!-- Eligible? Yes->Informed consent; No->Study stats collection -->
<sdm:Transition OID="TRANS.INCLUSION_EXCLUSION"
 Name="Transition from checking inclusion/exclusion criteria"
 SourceActivityOID="ACT.INCLUSION_EXCLUSION">
 <sdm:Switch>
  <sdm:TransitionDestination ConditionOID="COND.ELIGIBLE_PATIENT"
   Name="Transition from checking inclusion/exclusion to informed
   consent when patient is eligible"
   OID="TRANS.INCLUSION_EXCLUSION_TO_INFORMED_CONSENT"
   TargetActivityOID="ACT.INFORMED_CONSENT"/>
  <sdm:TransitionDefault Name="Transition from checking
   inclusion/exclusion to study stats collection when patient is
   not eligible"
   OID="TRANS.INCLUSION_EXCLUSION_TO_STUDY_STATS"
   TargetActivityOID="ACT.STUDY_STATS"/>
 </sdm:Switch>
</sdm:Transition>
```

Figure 3. Example Transition and Switch elements

Timing Constraints are defined in specific Timing section of the SDM-XML document and they determine constraints on activities and workflow transitions. Some of the core concepts of the timing constraint are now described, with reference to one example in the GERD study: The first visit PROM collection in the continuous arm should take place 1 day after the outcome of the participant randomization is known. It is allowed to occur 1 day before or 3 days after the target time.

```
<sdm:RelativeTimingConstraint
  Name="First visit to continuous arm first visit PROM timing constraint"
  OID="TCR.FIRST_VISIT_TO_FIRST_VISIT_CONTINUOUS_PROM"
  PredecessorActivityOID="ACT.RANDOMISATION_OUTCOME"
  SuccessorActivityOID="ACT.FIRST_VISIT_CONTINUOUS_PROM"
  TimepointRelativeTarget="P1D"
  TimepointGranularity="PD"
  TimepointPreWindow="P1D"
  TimepointPostWindow="P3D"
  Type="FinishToStart">
  <Description>
    <TranslatedText xml:lang="en">(Continuous arm) first visit PROM within
    3 days after first visit</TranslatedText>
  </Description>
</sdm:RelativeTimingConstraint>
```

Figure 4. Timing constraint for the first visit PROM collection in the continuous arm

Constraints can be absolute and relative. Absolute constraints limit when an activity can take place, while relative timing constraints specify the timing of two activities relative to one another. In Figure 4, the randomization outcome and the first visit PROM collection in the continuous arm are relative activities.

An ideal time for the elapsed time between activities and workflow transitions can be specified, as a timing window. In Figure 4, the pre-window duration is 1 day, while the post-window duration is 3 days. The time-point granularity option in days, referred to as "PD", means that the PROM collection can happen at any time in that day, and is still within 1 day after the end of the first visit, when the randomization outcome is known to the GP.

The timing type describes the relationship of one activity relative to another, e.g. FinishToStart shows that the PROM collection (subsequent activity) should start after the randomization outcome activity (preceding activity) is completed.

3.2. Study data collection model

The Operational Data Model (ODM) is a vendor-neutral, platform independent format for the interchange and archive of clinical study data [8]. The format is especially suitable when multiple systems and data sources are involved. The clinical research environment has existing clinical data management systems (CDMS); many already support the ODM format. In TRANSFoRm, the ODM format is being used for the interchange of clinical research data.

The ODM models the study data as consisting of several types of entities:

- **Item:** Individual clinical data item, such as weight measurement.
- **Item group:** Closely related items are collected together into item groups.
- **Form:** A form collects a set of logically and temporally related information and represents a page in a CRF on screen or on paper.
- **Study event:** A series of forms are collected as part of a study event.

A study protocol may have a number of planned visits by study participants, and each visit can correspond to one or more study events. In the ODM XML document, the metadata entities StudyEventDef, FormDef, ItemGroupDef and ItemDef describe the types of entities allowed in the study.

Relevant attributes that are used in the GERD study are described below and Figure 5 shows an extract from the study definition. The GERD study file is a snapshot and contains metadata:

- StudyOID: Clinical data entities can be identified with internal or external keys. The StudyOID uniquely identifies a study.
- FileType: This can be Snapshot (current state of the database with no information of the state over time) or Transactional (shows both current state and prior states of the database).
- Granularity: This indicates the breadth of information in the document, such as All (any type of data and metadata), Metadata (only metadata), AllClinicalData (clinical data only).

```
<?xml version="1.0" encoding="UTF-8"?>
<ODM xmlns="http://www.cdisc.org/ns/odm/v1.3"
     xmlns:sdm="http://www.cdisc.org/ns/studydesign/v1.0"
     xmlns:transform="http://www.transformproject.eu/v1.0"
     CreationDateTime="2013-10-28T14:05:28"
     Description="GORD Pilot Study Design, CDISC ODM version 1.3 format"
     FileOID="GORD_PILOT"
     FileType="Snapshot"
     Granularity="Metadata"
     ODMVersion="1.3"
     SourceSystem="TRANSFoRm"
     SourceSystemVersion="1.0">
  <Study OID="GORD-Pilot">
```

Figure 5. ODM file attributes for the GERD study

Metadata elements for the GERD study are:

StudyEventDef: Form(s) that are part of scheduled events, unscheduled events (e.g. study termination due to serious adverse event) or common study events (e.g. adverse events). Figure 6 shows the definition for the unscheduled visits of participants in the on-demand arm, when CROMs need to be collected. The Repeating attribute indicates that this event can occur repeatedly for the participant.

```
<!-- Study event: on-demand arm event visit -->
<StudyEventDef  Name="On-demand Event Visit"
                OID="SE.ONDEMAND_EVENT_VISIT"
                Repeating="Yes"
                Type="Unscheduled">
  <Description>
    <TranslatedText xml:lang="en">On-demand Arm Event Visit
    </TranslatedText>
  </Description>
  <FormRef  FormOID="FD.EVENT_VISIT_CROM"
            OrderNumber="1"
            Mandatory="Yes"/>
  <sdm:ActivityRef  ActivityOID="ACT.EVENT_VISIT_ONDEMAND_CROM"
                    OrderNumber="1"/>
</StudyEventDef>
```

Figure 6. StudyEventDef element example

FormDef: Describes a form that is used in a study. For example, Figure 7 shows the definition of the

Event Visit CROM form, which was referenced in Figure 6. This form does not occur repeatedly within the study event, as indicated by the Repeating attribute. This form has references to 13 item groups.

```
<!-- Event visit CROM -->
<FormDef  OID="FD.EVENT_VISIT_CROM" Name="Event Visit CROM Form"
          Repeating="No">
  <Description>
    <TranslatedText xml:lang="en">Event Visit CROM</TranslatedText>
  </Description>
  <!-- ID -->
  <!-- Date -->
  <ItemGroupRef ItemGroupOID="IG.VISIT_CAUSE" Mandatory="Yes"/>
  <ItemGroupRef ItemGroupOID="IG.WEIGHT" Mandatory="Yes"/>
  <ItemGroupRef ItemGroupOID="IG.LW_PPI" Mandatory="Yes"/>
  <ItemGroupRef ItemGroupOID="IG.LW_OTHER_REFLUX" Mandatory="Yes"/>
  <ItemGroupRef ItemGroupOID="IG.PRESENT_DIAGNOSIS" Mandatory="Yes"/>
  <ItemGroupRef ItemGroupOID="IG.ENDOSCOPY_LV" Mandatory="Yes"/>
  <ItemGroupRef ItemGroupOID="IG.ESOPHAGITIS" Mandatory="Yes"/>
  <ItemGroupRef ItemGroupOID="IG.BARRETTS" Mandatory="Yes"/>
  <ItemGroupRef ItemGroupOID="IG.OESOPHAGEUS_OTHER_CHANGES"
  Mandatory="Yes"/>
  <ItemGroupRef ItemGroupOID="IG.SICK_LEAVE_LV" Mandatory="Yes"/>
  <ItemGroupRef ItemGroupOID="IG.SMOKE" Mandatory="Yes"/>
  <ItemGroupRef ItemGroupOID="IG.CROM_ALARM_SYMPTOM" Mandatory="Yes"
  />
  <ItemGroupRef ItemGroupOID="IG.HEALTH_STATUS" Mandatory="Yes"/>
  <transform:FormType>CROM</transform:FormType>
</FormDef>
```

Figure 7. FormRef element for an event visit CROM

ItemGroupDef: Describes an item group that can occur in a study. Figure 8 shows the weight item group definition. It is a non-repeating group that has 2 items: the weight value and the weight unit, referenced by the ItemRef element.

```
<!-- Weight -->
<ItemGroupDef OID="IG.WEIGHT" Name="Weight ItemGroup" Repeating="No">
  <Description>
    <TranslatedText xml:lang="en">Weight (kg)</TranslatedText>
  </Description>
  <ItemRef OrderNumber="1" ItemOID="ID.WEIGHT" Mandatory="Yes"/>
  <ItemRef OrderNumber="2" ItemOID="ID.WEIGHT_UNIT" Mandatory="Yes"/>
  <transform:Query>fc91f02c-4dff-428e-ace7-1ad35a7b0093
  </transform:Query>
</ItemGroupDef>
```

Figure 8. ItemGroupDef element for weight

ItemDef: A type of item that occurs within a study. It can have properties, such as data type, range, and codelist restrictions. Figure 9 shows the definition of the weight value item. It is a floating point number with 2 decimal points. The Question element shows the text that prompts a user to provide the data for this item. The RangeCheck element constrains the value to > 0, with an ErrorMessage element to display an error when the range check is violated. The transform:CdimBinding element at the end of ItemDef (line 15) will be described with other semantic extensions to ODM. This links the weight item to the ontology of clinical primary care domain.

```
<ItemDef  OID="ID.WEIGHT"
          Name="Weight Item"
          DataType="float"
          Length="5"
          SignificantDigits="2">
  <Question>
    <TranslatedText xml:lang="en">Weight</TranslatedText>
  </Question>
  <RangeCheck SoftHard="Hard" Comparator="GT">
    <CheckValue>0</CheckValue>
    <ErrorMessage>
      <TranslatedText xml:lang="en">Weight must be more than 0.
      </TranslatedText>
    </ErrorMessage>
  </RangeCheck>
  <transform:CdimBinding>CDIM_000068</transform:CdimBinding>
</ItemDef>
```

Figure 9. ItemDef element for the weight item

3.3. Randomization

Study participants will be randomized to either on-demand or continuous use of PPI for 8 weeks. Patients randomized to on-demand use will be prescribed PPI in the presence of symptoms but no more than 40 mg omeprazole per day. Patients randomized to continuous use of PPI will be prescribed 20 mg omeprazole per day. This will be a block randomization containing 4 blocks based on age (50 years and below, and above 50 years) and gender. The study is not blinded so there is no need to break the randomization code.

The input parameters to the block randomization function (i.e. age and gender) are captured by the Randomization Form, which will be triggered by the randomization activity at patient's first visit. Collected data are sent to the TRANSFoRm Study System, which assigns the patient to a block and applies the randomization function. The outcome is hidden from the GP until the end of the consultation when the dose is prescribed. The outcome is then informed through the Randomization Outcome Form.

The TRANSFoRm randomization function requires proper configuration of the blocking variables in the study definition. In the case of the GERD study, the blocking variables are year of birth and gender. Because the current version of SDM XML does not cover randomization design, the blocking variables have to be specified in a TRANSFoRm extension element <transform:ParticipantSegmentVariables>. As shown in Figure 10, the variables year of birth and gender are linked to the corresponding fields in the Randomization Form through the ItemOID, and their range definitions specify the group arrangement. Another parameter required by the randomization function is the estimated number of recruited participants. The number is used to initialize the random allocation sequence. This parameter is defined as a <sdm:Parameter> in the <sdm:Summary> section.

```
<transform:ParticipantSegmentVariables xmlns="
http://www.transformproject.eu/v1.0">
  <Variable itemRefOID="ID.BRTHYR" varOID="PSV.Var1">
    <Name>Year of Birth</Name>
    <Description>
      <TranslatedText xml:lang="en">Year of birth of patients
      </TranslatedText>
    </Description>
    <Type>integer</Type>
    <Unit/>
    <Ranges>
      <Range rangeOID="Var1.R1">
        <Relation>GE</Relation>
        <Value>1964</Value>
      </Range>
      <Range rangeOID="Var1.R2">
        <Relation>LT</Relation>
        <Value>1964</Value>
      </Range>
    </Ranges>
  </Variable>
  <Variable itemRefOID="ID.GENDER" varOID="PSV.Var2">
    <Name>Gender</Name>
    <Description>
      <TranslatedText xml:lang="en">Gender of patients
      </TranslatedText>
    </Description>
    <Type>text</Type>
    <Unit/>
    <Ranges>
      <Range rangeOID="Var2.R1">
        <Relation>EQ</Relation>
        <Value>MALE</Value>
      </Range>
      <Range rangeOID="Var2.R2">
        <Relation>EQ</Relation>
        <Value>FEMALE</Value>
      </Range>
    </Ranges>
  </Variable>
</transform:ParticipantSegmentVariables>
```

Figure 10. Randomization blocking definition

4. Detailed Clinical Modeling Approach and Archetypes

In the previous section, we described a model-based approach to representing clinical research studies. TRANSFoRm achieves the reuse of routinely collected clinical data will be achieved by pre-populating eCRFs with available clinical data directly from EHR systems. Thus, the clinical data needs to be semantically interoperable between the eCRF (TRANSFoRm study system) and the EHR system. We adopt a two-level modelling approach [1,25] to separate out the more stable domain information from the various schemata implemented by the heterogeneous data sources [21].

Detailed Clinical Models (DCM) organize health information by combining knowledge, data element specification, relationships between elements, and terminology into information models that allow deployment in different technical formats [14,15]. DCM enables semantic interoperability by formalizing or standardizing clinical data elements, which are modeled 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.

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) [20], which defines the workflow and data requirements of the clinical research task, combined with the Clinical Data Integration Model (CDIM) [11,12], 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 [27]. 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 modeling approach, using the concept of archetype for detailed clinical content modeling, has been adopted by ISO/CEN 13606 [22]. Based on the OpenEHR framework [24], 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.

4.1. Pre-populating eCRFs from EHR

Table 1 lists the data elements, their archetypes and the visits (marked by 'X') at which they will be collected. Some data items are collected only at the time of first visit, while some others have to be collected every time. The data elements are listed together with their identifying terminology codes, where prescriptions use ATC codes and diagnosis use ICD-10 codes. Data extraction queries for retrieving these data elements are formulated as a DataExtractionQueryRequest for each (see Figure 11).

Table 1. EHR data elements for CROM form pre-population

Data Element	Archetype	Visit1	Visit2	Event Visit
Birth year (YYYY)	Date of Birth	X		
Gender	Gender	X		
Height	Height	X		
Weight	Weight	X	X	X
PPI use (A02BC)	Prescription	X		
Antacids (A02A) /	Prescription	X	X	X

alginate (A02BX13)				
Esophagitis (K20)	Diagnosis	X	X	X
Barrett's oesophagus (K22.7)	Diagnosis		X	X

4.2. Semantic Extension to ODM

A study participant's clinical data space is represented by a generic reference model. The reference model is based on CRIM, which is based on the Biomedical Research Integrated Domain Group (BRIDG) model of the semantics of protocol-driven clinical research [3]. Data elements retrieved from EHR sources are considered as PerformedObservationResult class (BRIDG class), extended so its attribute is of type Element (openEHR class). The tabular structure of PerformedObservationResult - Element is compatible with CDISC SDTM and ODM forms. Figure 11 shows how ODM can be semantically extended via the ItemGroupDef and ItemDef elements to be interoperable with EHR clinical data that are represented in CRIM reference model, with a binding to the CDIM ontology.

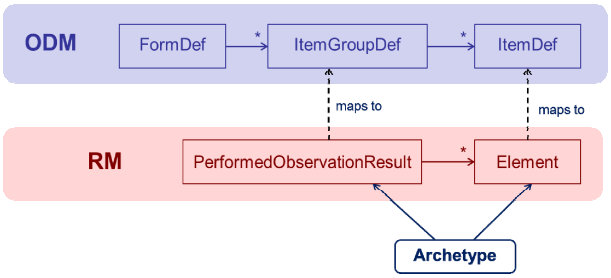

Figure 11. Link between ODM and the Reference Model via archetypes

For example, in Figure 9, the transform:CdimBinding element is a TRANSFoRm extension to ODM, for semantic interoperability. The weight concept is identified in the CDIM ontology as concept CDIM_000068. This binding maintains the link between the form item and the weight value after it is retrieved from the EHR system for pre-population.

In Figure 8, the last element in the item group definition for Weight is the transform:Query element, also a TRANSFoRm extension to ODM. This query ID refers to a data extraction request for weight value, unit and time of measurement, as shown in Figure 12, indicated by the three CdimConcept elements.

```
<!-- Weight -->
<DataExtractionRequest xmlns="http://www.transformproject.eu/query">
    <QueryId>fc91f02c-4dff-428e-ace7-1ad35a7b0093</QueryId>
    <ExtractQuery>
        <Select>
            <CdimConcept>CDIM_000068</CdimConcept>
            <CdimConcept>CDIM_000100</CdimConcept>
            <CdimConcept>CDIM_000067</CdimConcept>
        </Select>
        <Where id="3402">
            <Expression xsi:type="ns2:ArchetypeExpression"
                        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
                <Archetype>
                  archetype (adl_version=1.4)
                  TRANSFoRm-CRIM-EVENT.weight.v1
```

Figure 12. Partial data extraction query for weight

The request for data extraction of the three weight-related data items is accompanied by constraints specifying how to filter the data. These are specified in the Archetype section of the query, as shown in Figure 13 (definition), Figure 14 (ontology term definitions) and Figure 15 (ontology term bindings).

```
<Archetype>
  archetype (adl_version=1.4)
    TRANSFoRm-CRIM-EVENT.weight.v1

  concept
    [at0000]

  language
    original_language = &lt;[ISO_639-1::en]&gt;

  definition
    EVENT[at0000] matches {
      attributes cardinality matches {1..3; ordered} matches {
        ATTRIBUTE[at0001] matches {
          value matches {
            C_DV_QUANTITY &lt;
                          &gt;
          }
        }
        ATTRIBUTE[at0002] matches {
          value matches {*}
        }
        ATTRIBUTE[at0003] matches {
          value matches {*}
        }
      }
    }
```

Figure 13. Weight archetype: definition

```
ontology
  terminologies_available = &lt;"CDIM", ...&gt;
  term_definitions = &lt;
    ["en"] = &lt;
      items = &lt;
        ["at0000"] = &lt;
          text = &lt;"Human Weight"&gt;
          type = &lt;"Event"&gt;
        &gt;
        ["at0001"] = &lt;
          text = &lt;"human weight data value"&gt;
          type = &lt;"Number"&gt;
        &gt;
        ["at0002"] = &lt;
          text = &lt;"human weight units"&gt;
          type = &lt;"Unit"&gt;
        &gt;
        ["at0003"] = &lt;
          text = &lt;"human weight recording time"&gt;
          type = &lt;"Date"&gt;
        &gt;
      &gt;&gt;&gt;
```

Figure 14. Weight archetype: ontology definitions

```
term_binding = &lt;
  ["CDIM"] = &lt;
    items = &lt;
      ["at0001"] = &lt;[CDIM::CDIM_000068]&gt;
      ["at0002"] = &lt;[CDIM::CDIM_000100]&gt;
      ["at0003"] = &lt;[CDIM::CDIM_000067]&gt;
    &gt;&gt;&gt;
```

Figure 15. Weight archetype: ontology term bindings

The archetype approach establishes the semantic foundation for describing clinical data elements, and enables the semantic interoperability across different EHR systems. Based on the archetypes, a query model is developed to facilitate query formulation for eligible patient identification and patient data extraction. Furthermore, complex Boolean logic and temporal constraints can be constructed in the query criteria. The actual queries formulated by the model are encoded in XML format.

5. Discussion

Clinical research is struggling to achieve sufficient quantity and quality of available evidence. High cost and complexity of conducting research is reducing the numbers of clinicians and patients willing to participate in clinical trials.

Integration of clinical research studies into routine clinical practice has potential to bring a fundamental shift to the conduct of medical research. Facilitating subject identification and recruitment alone would significantly reduce the cost and complexity of running trials. Feeding EHR data into eCRF forms reduces effort for clinicians and decreases the likelihood of manual error occurring. Finally, consistency between the vocabularies and terminologies used in research and practice facilitates storing of collected research data into EHR systems for clinician's future reference.

Any such impactful change brings with it several implications and challenges. A methodological issue that this work raises is the strategy for defining local mapping models that TRANSFoRm mediation ontology, CDIM, uses. One approach is for it to be done centrally, by the development team, which was indeed how the GERD study was performed. However, another, more sustainable, strategy is to develop mapping tools that local experts can use to implement their own mappings. The advantage of this is that any local peculiarities, e.g. custom codes added to standard vocabularies, can be easily addressed by the experts who are familiar with them. These tools are in development by the TRANSFoRm team, and will be prototyped in further integrations.

On a conceptual level, our work establishes a standard-based method for connecting clinical research study systems and EHR repositories. This has been an area of interest for a number of research bodies. Particularly related to this work are CDISC-IHE Healthcare Link profiles [6], integration standards for connecting healthcare and clinical research, particularly the Retrieve Form for Data Capture Profile (RFD) and the Clinical Research Document Profile (CRD). RFD provides a method for gathering data within a user's current application to meet the requirements of an external system, while the CRD describes the content pertinent to the clinical research use case required within the RFD pre-population parameter. These profiles focus on web services for retrieving forms and using HL7 Continuity of Care (CCD) format [16] for pre-population of eCRFs from external CDMS to display within EHR systems. TRANSFoRm adopted an alternative strategy, by which eCRFs are deployed from within the EHR system itself. From our experience, EHR vendors in Europe are still in the early stage of supporting HL7 CCD and IHE profiles. On the other hand, these standardisation efforts do not yet offer the necessary semantic linkage required for lossless data sharing across systems.

Our LHS approach embeds clinical research into clinical practice. This is a *clinical study-centric activity*. The TRANSFoRm semantic interoperability framework will also be used to put research data back into the EHR to enrich the patient record, thereby completing the cycle. This *EHR-centric activity* now requires further investigation into how EHRs can best represent research data for future reuse. Most EHR systems support free-text data fields to be entered, but to realize the full benefit, a structured data storage is required, with appropriate coding to facilitate use in clinical practice.

6. Conclusion

This paper described the modeling aspects of the implementation of Learning Healthcare System that addresses the integration of clinical trials into the routine clinical practice, thereby reducing cost and complexity of the task. This is achieved through making a functional eCRF semantically interoperable with the EHR system to enable patient identification and pre-population of eCRFs with available primary clinical data. A two-level modeling approach has been adopted, with the first level being guided by the reference model for the clinical research domain (CRIM), and the second level using archetypes to constrain the reference model. Clinical data elements

have a binding to the CDIM ontology that describes the primary care clinical domain. To maintain the semantic meaning when using eCRFs, the CDISC ODM standard has been extended to allow for archetypes to further define form elements and queries to EHR systems to extract specific clinical data items.

The work has great potential to change the way clinical research is conducted, but it also opens up new methodological challenges. However, resolving these challenges will provide generic solutions that will not be restricted to individual health systems.

7. References

[1] T. Beale, "Archetypes: Constraint-based domain models for future-proof information systems," in 11th OOPSLA Workshop on Behavioral Semantics: Serving the Customer, pp. 16-32 Northeastern University 2002.

[2] D. Blumenthal, "Stimulating the adoption of health information technology," N. Engl. J. Med., vol. 360, no. 15, pp. 1477–1479, Apr. 2009.

[3] BRIDG Founding Stakeholders, "BRIDG Model." Available: www.bridgmodel.org/.

[4] R. M. Califf, "Clinical research sites—the underappreciated component of the clinical research system," JAMA, vol. 302, no. 18, pp. 2025–2027, Nov. 2009.

[5] Clinical Data Interchange Standards Consortium, CDISC. Available: www.cdisc.org/.

[6] CDISC, "CDISC Healthcare Link Profiles Information." Available: www.cdisc.org/healthcare-link.

[7] Clinical Data Interchange Standards Consortium, "Study/Trial Design Model: Study Design in XML Version 1.0." Available: www.cdisc.org/study-trial-design.

[8] Clinical Data Interchange Standards Consortium, "Operational Data Model." www.cdisc.org/odm.

[9] V. Curcin et al., "Implementing interoperable provenance in biomedical research," Future Gener. Comput. Syst., vol. 34, pp. 1–16, May 2014.

[10] EHR4CR Consortium, Electronic Health Records for Clinical Research. Available: www.ehr4cr.eu.

[11] J.-F. Ethier et al., "A unified structural/terminological interoperability framework based on LexEVS: application to TRANSFoRm," J. Am. Med. Inform. Assoc., Apr. 2013.

[12] J.-F. Ethier et al., "Clinical Data Integration Model: Core Interoperability Ontology for Research Using Primary Care Data (Accepted)," Methods Inf. Med., 2014.

[13] C. P. Friedman, A. K. Wong, and D. Blumenthal, "Achieving a nationwide learning health system," Sci. Transl. Med., vol. 2, no. 57, p. 57cm29, Nov. 2010.

[14] W. Goossen, A. Goossen-Baremans, and M. van der Zel, "Detailed Clinical Models: A Review," Healthc. Inform. Res., vol. 16, no. 4, pp. 201–214, Dec. 2010.

[15] W. T. F. Goossen and A. Goossen-Baremans, "Bridging the HL7 template - 13606 archetype gap with detailed clinical models," Stud. Health Technol. Inform., vol. 160, no. Pt 2, pp. 932–936, 2010.

[16] HL7, "HL7/ASTM Implementation Guide for CDA® R2 -Continuity of Care Document (CCD®) Release 1." www.hl7.org/implement/standards/product_brief.cfm?product_id=6.

[17] G. Hripcsak and D. J. Albers, "Next-generation phenotyping of electronic health records," JAMIA, vol. 20, no. 1, pp. 117–121, Jan. 2013.

[18] G. Jiang et al., "Harmonization of Detailed Clinical Models with Clinical Study Data Standard," in 2nd International Workshop on Managing Interoperability and complexity in Health Systems, New York, NY, USA, 2012, pp. 23–30.

[19] I. L. Katzan and R. A. Rudick, "Time to integrate clinical and research informatics," Sci. Transl. Med., vol. 4, no. 162, p. 162fs41, Nov. 2012.

[20] W. Kuchinke, T. Karakoyun, and C. Ohmann, "Clinical Research Information Model," Project Deliverable TRANSFoRm Deliverable D6.2, V1.0, May 2012.

[21] S. N. Lim Choi Keung et al., "Detailed Clinical Modelling Approach to Data Extraction from Heterogeneous Data Sources for Clinical Research," in 2014 Summit on Clinical Research Informatics, San Francisco, USA, 2014.

[22] P. Muñoz et al., "The ISO/EN 13606 Standard for the Interoperable Exchange of Electronic Health Records," J. Healthc. Eng., vol. 2, no. 1, pp. 1–24, 2011.

[23] L. Olsen, D. Aisner, and J. M. McGinnis, "The Learning Healthcare System: Workshop Summary (IOM Roundtable on Evidence-Based Medicine)," 2007. Available: www.nap.edu/openbook.php?record_id=11903.

[24] openEHR Foundation, openEHR. Available: www.openehr.org

[25] A. L. Rector et al., "A framework for modelling the electronic medical record," Methods Inf. Med., vol. 32, no. 2, pp. 109–119, Apr. 1993.

[26] SemanticHealthNet Consortium, Semantic Health Net. Available: www.semantichealthnet.eu.

[27] B. Smith and W. Ceusters, "HL7 RIM: an incoherent standard," Stud. Health Technol. Inform., vol. 124, pp. 133–138, 2006.

[28] N. S. Sung et al., "Central challenges facing the national clinical research enterprise," JAMA J. Am. Med. Assoc., vol. 289, no. 10, pp. 1278–1287, Mar. 2003.

[29] TRANSFoRm Consortium, Translational Research and Patient Safety in Europe. Available: www.transformproject.eu.

[30] T.-P. van Staa et al., "Pragmatic randomised trials using routine electronic health records: putting them to the test," BMJ, vol. 344, p. e55, 2012.

[31] V. Venkatesh et al., "User Acceptance of Information Technology: Toward a Unified View," MIS Q., vol. 27, no. 3, pp. 425–478, Sep. 2003.

[32] C. Weng et al., "An Integrated Model for Patient Care and Clinical Trials (IMPACT) to support clinical research visit scheduling workflow for future learning health systems," J.Biomed.Inform vol. 46, no. 4 pp.642–652, 2013.