Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)

doi:10.3233/978-1-61499-753-5-156

The 'PEARL' Data Warehouse: Initial Challenges Faced with Semantic and Syntactic Interoperability

Samhar MAHMOUD[a,1], Andy BOYD[b], Vasa CURCIN[a], Richard BACHE[a], Asad ALI[b], Simon MILES[a], Adel TAWEEL[a], Brendan DELANEY[c] and John MACLEOD[b]
[a] *King's College London*
[b] *University of Bristol*
[c] *Imperial College London*

Abstract. Data about patients are available from diverse sources, including those routinely collected as individuals interact with service providers, and those provided directly by individuals through surveys. Linking these data can lead to a more complete picture about the individual, to inform either care decision making or research investigations. However, post-linkage, differences in data recording systems and formats present barriers to achieving these aims. This paper describes an approach to combine linked GP records with study observations, and reports initial challenges related to semantic and syntactic interoperability issues.

Keywords. Data Linkage, Electronic Health Records, PEARL, ALSPAC

1. Introduction

Cohort studies increasingly implement comprehensive record linkage programs to retrospectively and prospectively collect information on study participants. To effectively combine linked data with survey data it is necessary to bring these disparate data into one single, heterogeneous representation. The Project to Enhance ALSPAC through Record Linkage (PEARL) is developing a data processing and warehousing solution (DWH) to help resolve these issues. The approach typically taken is to create a single DWH constructed according to a well-defined data model. In a clinical context, electronic patient records (EPRs) are typically arranged as an event sequence. In contrast, observational studies tend to collate observations in wide, flat, file structures; where each record represents a participant, and each file a data collection exercise. The PEARL DWH combines both clinical and self-reported information into a single event record. This format is familiar to clinical researchers, is novel in a cohort context, and contrasts with other contemporary approaches [1,2]. This format allows users to efficiently extract data using standard querying languages available in routine analytical software. This paper describes: the chosen data model; data pipeline workflows that combine EPRs and self-reported participants' data of the Avon Longitudinal Study of Parents and Children (ALSPAC); and, the methods developed to overcome interoperability challenges.

[1] Corresponding author, Division of Health & Social Care Research, Faculty of Life Sciences & Medicine, King's College London, The Strand, London, WC2R 2LS; E-mail: samhar.mahmoud@kcl.ac.uk.

2. Background

ALSPAC is a longitudinal birth cohort study collecting information of participants' life course exposures, and health, social and well-being outcomes. ALSPAC recruited pregnant women - living in, and around, the City of Bristol - due to deliver between 01/04/91 and 31/12/92 [3]. An initial total of 14,062 live-born children were enrolled. Data is collected via questionnaires, study assessment visits, biological and 'omic characterisations (see: www.bristol.ac.uk/alspac/researchers/access/). PEARL (PI John Macleod) was designed to complement these self-reported data through the secondary use of linked routinely collected records. Ethical approval was obtained from the ALSPAC Ethics and Law Committee and NHS Research Ethics Committee (Ref: Haydock 10/H1010/70). To help ensure acceptable data usage, PEARL implemented a 'Data Safe Haven' governance framework [4]. The safe haven incorporates a 'UK Secure eResearch Facility' (UKSeRP) developed by the Welsh Farr Institute as a secure data repository and analysis platform. In 2013 Boyd and Macleod, with Egton Medical Information Systems (EMIS) Ltd and Apollo Medical Systems (Apollo) Ltd, used this framework to extract a pilot 3,166 EPR instances, relating to 2,249 ALSPAC participants, from 181 General Practices. An exemplar use-case was identified to test the functionality of the data warehouse. The use case – an investigation into genetic and environmental influences on asthma – will use linked EPRs to assess: i) clinical validation of self-reported data; ii) impact of prescribed treatment; and, iii) value of EPR data in missing data methods. The initial data migrated into the data warehouse relate to this use case.

3. Method

Within clinical data, we can distinguish between existential facts and value-bearing facts. Existential facts record only the fact that something occurred e.g. a diagnosis. Value-bearing facts record not just an occurrence but also some value associated with it such as a BMI reading and the corresponding value and unit of measurement. Three types of structured value have been identified in both clinical and non-clinical data and we use the ISO21090 data types [5], which are: Physical quantity (ISO21090 type PQ) where there is a numeric (scalar) quantity and a unit of measurement; Coded Ordinal (ISO21090 type CO) where there is an ordinal scale of values to which both a numeric value and a meaning (a code) are assigned; and Coded Value (ISO21090 type CD) where there is no numeric value or order over a range of values.

An event-based model is chosen, since we believe that all data sources can be rendered into a sequence of events. Thus, each data point is recorded as a distinct fact, with a single timestamp relates to precisely one individual. GP data sources already use an event-based representation. Equally, a self-reported questionnaire can be viewed as a series of events where each answer to a question is a single event. Some existing, and widely used, standards define conventions for the exchange of healthcare data between clinical sites. Four of these adopt an event-based representation: HL7 Reference Information Model (RIM, www.hl7.org); OpenEHR (www.openehr.org); Continuity of Care Record (CCR); and Continuity of Care Document (CCD). CCD and CCR have significant limitations for longitudinal studies, in that they do not support non-numeric values (CO and CD) - such as those generated by most questionnaire questions - and that for PQ data elements only a single reading is actually stored. Both OpenEHR and HL7 RIM support all the above data types. However, the availability of a database schema

with an inbuilt provenance model made the HL7 option the obvious one. We have also adopted the ISO 21090 data types known as the Constrained Information Model (CIM) [5], which has not, to date, been used to store non-clinical data.

3.1. The HL7 RIM-based Data Model

The main classes in the data model are, as shown in Figure 1: Subject – contains details of birth, death and administrative gender; Clinical Statement – records events such as: Observations – diagnoses, reporting of symptoms, measurements and any other observation, and Procedures – anything performed on the subject, of which a special case is Substance administration – medications, vaccines etc; and Organisation.

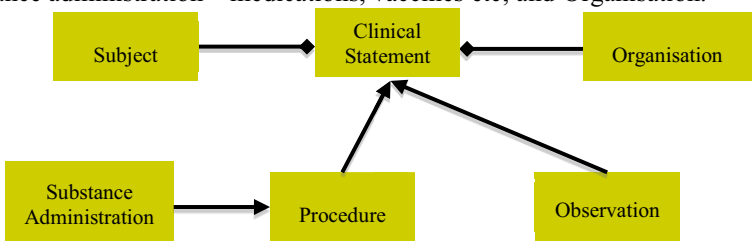

Figure 1. High-level class diagram of data model

For observations, there are 'code' and 'value' fields. For existential observations, the 'code' field is set to a coded value indicating a 'diagnosis' or 'finding'. The 'value' field is used to record the nature of the diagnosis. For value-bearing observations, the 'code' is again set to a coded value (e.g. indicating a laboratory test or vital sign). The 'value' field is used to record the reading. Here, the value may be of data type PQ, CO and CD. For data type CO, both a code and an (ordinal) numeric value are stored. For data type PQ a (scalar) value and unit of measurement are recorded. HL7 requires that all units of measurement must be recorded in UCUM (Universal Convention for Units of Measurement, www.dmd.nhs.uk). HL7 also requires that where a coding system is used it should be specified by means of its OID (Object Identifier). Therefore, each concept is specified by a code and OID defining the coding system and optionally a display name describing the concept and a free text string describing the coding system.

Table 1. Summary of GP Data from Two Sources

Attribute	EMIS Data	Apollo Data
No. Patient Record Instances	2787	379
No. Clinical Facts	400975	60668
No. Clinical Facts Processed	377681	53476
Non-medication Coding system	Read v2, SNOMED-UK	Read v2
Medication Coding System	SNOMED-UK, Local codes	Read v2
Units of Measurement	Mostly present	Missing
Format	OpenHR XML	CSV files

A majority of data (EMIS) is exported as XML files according to OpenEHR schema. A minority of the data (Apollo) is held in tab-delimited files. Table 1 summarises key features of the two data sources. Units of measurement, which are necessary for PQ-value facts, have two problems. First, it is missing in Apollo data and for some of EMIS data. Second, in EMIS data, the units are not specified in UCUM. Thus two further mappings are required: Non-standard unit to UCUM unit; and Read v2 codes for vital sign, lab test etc. and actual value to the UCUM unit. These were constructed on an ad hoc basis by PEARL staff and confirmed by clinical review.

3.2. Incorporating Questionnaire Data into HL7 RIM

Non-clinical data may relate to medical matters, such as self-reported (or parent/ carer) disease, symptom or medication use. It may also relate to non-medical matters such as housing, diet, lifestyle or opinions (e.g. do you like your teacher?). Of the non-clinical data available to PEARL, only questionnaire data has been formally considered to date. The Asthma use-case comprised of a subset of 217 questions from 34 ALSPAC mother and child questionnaires. All questions from the use case required a constrained response, where participants had to complete either a multiple-choice answer, with a prescribed set of answers or a numerical or timestamp reply.

Non-clinical data are recorded as observations, which do not differ substantially from clinicians recording survey responses, or consultation questions without any independent corroboration. Therefore, non-clinical questionnaire data with structured replies can be recorded as a value-bearing observation. Analysis of the asthma use case shows that all replies can be recorded as data types PQ, CD and CO.

The terms describing clinical facts are comprehensively recorded within structured coding systems (e.g. Read or SNOMED codes). However, survey responses do not have comprehensive code frames (although domain specific vocabularies' such as HASSET and MESH exist). Therefore, we have developed KASPER (King's Auxiliary System for Provisionally Encoding Records). PEARL uses SNOMED codes, where they exist, to code attributes such as height or body mass. Where questionnaire concepts are not supported by any terminology, a new KASPER code is created for the *root* concept. For data of types CO and CD, KASPER leaf codes are created for each of the possible replies that can be given. KASPER has its own OID registered with HL7. Figure 2 gives an example of how replies to questions are mapped to KASPER codes. The PEARL data is supplied with each subject being identified by the UKSeRP anonymised linkage field (ALF). This is unique to each subject and is the means by which records from different sources can be linked within the UKSeRP environment.

Code	Display Name	Value (code)	Value (number)	Display Name
S1003R	Disinfectant exposure in utero	S1002L3	3	Once per week
S1004R	Bleach exposure in utero	S1002L2	2	Most days
S1005R	Window cleaner exposure in utero	S1002L4	4	< Once per week

Figure 2. Mapping of Questionnaire Replies to KASPER codes

Questionnaires results are held in a CSV files. Each row represents one subject. The meaning of each variable is defined in the ALSPAC documentation. The data model requires that each attribute addressed either by a questionnaire question or composite score be represented by a code. Each variable in the asthma use case was checked against the ALSPAC documentation. Where the attribute had an existing SNOMED code e.g. body height, this was used. Otherwise a KASPER code was created. For CD and CO data types, it is necessary to create also leaf codes for each possible value. For variables of type PQ a unit of measurement in UCUM was specified.

3.3. Data Pipelines

The EPR extract and loading process has three stages. 1) Clinical facts are extracted and parsed into objects of type 'Fact'. 2) Data are cleaned and transformed to create a 'PEARL Fact' object, where: a 'Fact' with numerical values is set to a PQ-value observation; a 'Fact' with a code listed in the CD or CO mappings is set to a CD-value

or CO-value observation respectively; a 'Fact' with a code listed in the vaccine mappings is assumed to be a substance administration. 3) 'PEARL Facts' are loaded into the DWH.

4. Conclusion

There is a great importance for the linkage of clinical data to make it available for medical research. This paper describes an approach utilising an event based data model to combine participants' clinical data with their self-reported questionnaire data. A number of broad approaches exist, but mainly focus on clinical data of the same nature [7], or choose alternate data models [1,2]. The primary challenges in adopting event based models relate to syntactical issues with the representation of questionnaire data. This was addressed through time-consuming clerical work, which is likely to be unsustainable over the whole of the ALSPAC data set or in other studies. In future, alternative coding schemes – such as MESH, HASSET or the Data Documentation Initiative (DDI) formatted cohort data dictionary developed by the CLOSER cohort consortium (http://www.closer.ac.uk/data-resources/closer-search-platform/) may be more efficient if they are found to fully combine clinical and inter-disciplinary survey domains. Other problems relate to missing clinical measurement units, which can be resolved by changing the EPR extraction protocols. Despite some initial problems, our work to date suggests the validity of this data warehouse design. Future work will integrate a wider range of data (ALSPAC clinical assessment data, genetic data) to complete our use case data set and then use the data warehouse to conduct exemplar epidemiological investigations using our asthma use case.

Acknowledgements

We are extremely grateful to the ALSPAC families, recruiting midwives and the ALSPAC team. The UK Medical Research Council, Wellcome Trust (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. The authors will serve as guarantors for the contents of this paper. This research was specifically funded by The Wellcome Trust (Grant ref: WT086118/Z/08/Z).

References

[1] Denazas D, George J, Herrett, et al. Data Resource Profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *Int J Epidemiol* (2012) 41 (6): 1625-1638.
[2] Ford DV, Jones KH, Verplancke J-P, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* (2009) 9:157.
[3] Boyd A, Golding J, Macleod J, et al. Cohort profile: the 'children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol*. 2012 Apr 16:dys064.
[4] Burton PR, Murtagh MJ, Boyd A, et al. Data Safe Havens in health research and healthcare. *Bioinformatics*. 2015;31(20):3241–8. doi:10.1093/bioinformatics/btv279.
[5] Bache R, Daniel C, James J, et al. An Approach for Utilizing Clinical Statements in HL7 RIM to Evaluate Eligibility Criteria, accepted for 25th European Medical Informatics Conference - MIE2014.
[6] ISO 21090:2011 Health informatics -- Harmonized data types for information interchange, International Standards Organisation (ISO), 2011.
[7] Karmen C, Ganzinger M, Kohl CD, et al. A framework for integrating heterogeneous clinical data for a disease area into a central data warehouse. *Stud Health Technol Inform*. 2014;205 1060-1064.