An Eligibility Criteria Query Language for Heterogeneous Data Warehouses*

R. Bache[1, 2]; A. Taweel[1, 2]; S. Miles[1]; B. C. Delaney[2]

[1]Department of Informatics, King's College London, London, UK;
[2]Department of Primary Care and Public Health Sciences, King's College London, London, UK

Keywords
Eligibility criteria, domain language, electronic healthcare records

Summary

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

Objectives: The increasing availability of electronic clinical data provides great potential for finding eligible patients for clinical research. However, data heterogeneity makes it difficult for clinical researchers to interrogate sources consistently. Existing standard query languages are often not sufficient to query across diverse representations. Thus, a higher-level domain language is needed so that queries become data-representation agnostic. To this end, we define a clinician-readable computational language for querying whether patients meet eligibility criteria (ECs) from clinical trials. This language is capable of implementing the temporal semantics required by many ECs, and can be automatically evaluated on heterogeneous data sources.

Methods: By reference to standards and examples of existing ECs, a clinician-readable query language was developed. Using a model-based approach, it was implemented to transform captured ECs into queries that interrogate heterogeneous data warehouses. The query language was evaluated on two types of data sources, each different in structure and content.

Results: The query language abstracts the level of expressivity so that researchers construct their ECs with no prior knowledge of the data sources. It was evaluated on two types of semantically and structurally diverse data warehouses. This query language is now used to express ECs in the EHR4CR project. A survey shows that it was perceived by the majority of users to be useful, easy to understand and unambiguous.

Discussion: An EC-specific language enables clinical researchers to express their ECs as a query such that the user is isolated from complexities of different heterogeneous clinical data sets. More generally, the approach demonstrates that a domain query language has potential for overcoming the problems of semantic interoperability and is applicable where the nature of the queries is well understood and the data is conceptually similar but in different representations.

Conclusions: Our language provides a strong basis for use across different clinical domains for expressing ECs by overcoming the heterogeneous nature of electronic clinical data whilst maintaining semantic consistency. It is readily comprehensible by target users. This demonstrates that a domain query language can be both usable and interoperable.

Correspondence to:
R. Bache
Dept. of Primary Care and Public Health Sciences
Floor 7, Capital House
Weston Street
London SE1 3QD
UK
E-mail: richard.bache@kcl.ac.uk

Methods Inf Med 2015; 54: 41–44
http://dx.doi.org/10.3414/ME13-02-0027
received: June 15, 2013
accepted: May 7, 2014
epub ahead of print: July 2, 2014

* Supplementary material published on our website www.methods-online.com

1. Introduction

Clinical data routinely collected for the care of patients is often used for research. Ideally, data from multiple sites would be accessed using a common federated query, for example in recruiting for clinical trials, where a researcher wishes to query clinical sites to obtain counts of patients meeting eligibility criteria (ECs), before identifying patients for recruitment.

However, creating a common federated query using a standard query language (e.g. SQL) to access different patient databases is severely limiting: i) it requires that all data warehouses have identical schemata and semantic representation; ii) the semantic power of such query languages is restricted, specifically where temporal semantics are required, as is often the case in ECs [1]; iii) it requires the researchers have technical knowledge not only of the terminology but also the database schema. As discussed in Section 3.1, existing attempts to express ECs in a formal way are not suitable for querying such databases.

We propose, therefore, to define a domain query language that uses a clinician-readable notation for accessing heterogeneous databases containing patient data. It should be powerful enough to express ECs that can be evaluated automatically from patient data, including the temporal constraints often found in ECs, such as finding the most recent test result or requiring that one event occurring after another. It should be implementable to access multiple database structures and semantic representations.

In this paper, we explain the design and implementation of such a language, ECLECTIC (Eligibility Criteria Language for Clinical Trial Investigation and Con-

struction). ECLECTIC is used in the Electronic Health Records for Clinical Research project (EHR4CR) [2], which aims to facilitate feasibility studies and recruitment of eligible patients by querying clinical data warehouses located at multiple sites across the EU.

2. Objectives

We propose a computable query language to express ECs for a clinical trial that can be evaluated from patient data without human intervention. It should be independent of specific semantic representations such as coding systems, noting that automatic conversions between alternative representations are often possible, and be independent of the structure of the data warehouses by instead referring to high-level concepts such as procedure or diagnosis, so its users would not need to know the data warehouse schemata.

This language is intended to express ECs that could be evaluated precisely from the structured data found at clinical sites; natural language processing of free-text fields is beyond the scope of this paper. It should be able to impose conditions not only on the existence or value of patient attributes but also their timing, as around 47% of ECs contain some temporal condition [1].

The language should be able to represent ECs in a clinician-readable way whilst also being sufficiently precise to express clinical concepts, physical quantities and time. It is not our intention to design a language for formalising free-text ECs generally as has been attempted elsewhere [3], but to define one that can express ECs and use them to automatically query diverse data warehouses.

3. Methods

We analysed a set of ECs and determined their elements and constructs, and the types of clinical data available in a patient record to evaluate them. To generalise, as much as possible, we based our analysis on two standards, HL7 RIM [8] and Continuity of Care Record (CCR) [9] to determine the possible types of structured clinical events that would be recorded. We found HL7 more comprehensive in identifying clinical events for our approach, however CCR terminology was more clinician-readable. We then constructed a clinician-readable notation forming a semantically enabled formal language to express queries. This was translated into an object model that provides a computational representation of the queries, independent of the representation of data in any one clinical data warehouse. The language and model were implemented and used by researchers for translating eligibility queries into a form that could be executed on the databases.

3.1 Existing Notations and Standards

Other notations were evaluated to determine if these already met our needs. Many EC representations have been devised and a comprehensive survey by Weng et al [4] examines 27 of them. If we exclude notations that relate only to a medical specialism, e.g. cancer or HIV, those that assume a specific data representation and those that rely on informal natural language representations, this leaves just three [5–7].

Arden Syntax [5] and GELLO [6] are formal languages that may be used to express clinical reasoning. They employ a programming-language style of syntax and contain a range of operators, such as conditional branching and looping. This means that any implementation would require building a full compiler or interpreter for each separate database schema. Furthermore, the syntax of these languages could make them inaccessible for many users. ERGO [7] allows formal reasoning but uses a natural language representation of clinical concepts making any implementation problematic. Thus, no language met our need for a well-defined, domain-specific, expressive high-level representation for a rich range of ECs that is agnostic of the structure and semantic representation of the data sources, and could be understood by non-technical users.

Any language to express ECs will require a set of high-level clinical concepts such as diagnosis, lab test and medication. Various standards exist to provide definitions of such concepts, and we drew from HL7 RIM and CCR to conform to standard-based semantic annotations. We note other standards such as the Quality Data Model [10] provide similar but subtly different concept categories and, whilst these standards remain at variance, conformance to all is not possible.

3.2 Defining the Language

From HL7 RIM [8], we identified the following types of clinical events that would apply to the studied ECs: living subject (for demographic data), procedures, substance administration (for medications) and observations such as diagnoses, vital signs, lab tests and assessments in some objective form. All clinical events are associated with at least one recorded time.

An analysis of 123 criteria from eight public clinical trials was initially used to determine the expressiveness the language would require, summarised in ▶Table 1 (online appendix). In HL7, *structured* clinical events are stored as terms (codes), numeric and coded values. In order to evaluate a patient's eligibility for one criterion, a predicate needs to be evaluated as either true or false based on the clinical facts in the patient record. Therefore, three types of predicate were identified as sufficient:

- *existential* – whether something happened e.g. a diagnosis or procedure (as a term),
- *coded value* – compared to a finite number of non-numeric values e.g. gender,
- *numeric value* – compared to a range of scalars, usually with a unit of measurement e.g. blood sugar as a lab test value.

For numeric predicates, the value expressed in the clinical fact is compared directly to the defined range; arithmetic operators were not found to be needed and are thus not supported. ▶Figure 1 gives an example of ECLECTIC. Rules 2 and 3 give examples of existential predicates. Rule 1 has a coded-value predicate and rule 4 has a numeric valued predicate. An analysis of nine further trials with 85 criteria, provided by Pharma within the EHR4CR project, showed the predicate types identified

above to be sufficient for automatic evaluation of ECs.

For the expression of time, we followed Nigrin and Kohane [11] in considering all clinical facts as events with zero duration, with temporal constraints referring to events being *before* or *after* other events, or being the *first* or *last* event of a given kind. Where *before* and *after* operators are used, a temporal constraint links a rule to a temporal anchor of which three types were defined: birth date, a previous rule (see rule 3 in ▶Figure 1) or *now* (the time when the query is launched, used in rule 4). Since gender assignment is deemed a unique event, no *first* or *last* is needed (rule 1), whilst age would be expressed as the number of years after the birth event. We considered that comparing whether events happened at exactly the same time (*equals*) to the nearest second would not be useful.

ECs can be either inclusion or exclusion criteria, so *not* was used to mark the latter (rule 4), and can be combined with operators *and* and *or*. By adopting the convention of conjunctive normal form, the need for brackets is avoided.

Clinical concepts (e.g. lab tests, medications, procedures etc.) must be defined by a code within a coding system. Thus, a triple, consisting of a code, coding system name and human-readable display name, was used to define each concept and ensure well-defined clinical semantics. In rule 2, the code 'E11' in ICD-10 specifies diabetes mellitus type II. ECLECTIC is coding system agnostic, however, in the EHR4CR project various coding systems such as SNOMED-CT, ICD-10, LOINC and ATC have been employed.

To express high-level clinical concepts such as lab test or medications, we used CCR since this provides clinician-friendly concepts. The HL7 RIM classes, such as 'Observation' operate at a coarser level of granularity and therefore could not distinguish between a diagnosis, test result or clinical assessment. Each rule in ECLECTIC is based on one clinical event and characterised by exactly one high-level concept such as diagnosis (*problem* in CCR) or result, of which there are currently 11, plus *deceased* which is not part of CCR but was found necessary. For an illustration of ECLECTIC, an example for a clinical trial is included here (▶online appendix).

3.3 Executing Queries on Heterogeneous Data Sources

The evaluation of an ECLECTIC query on each data warehouse is performed in three stages: i) query specifications are generated which determine what data is required from the warehouse to evaluate the rules against; ii) a warehouse-specific adaptor maps these specifications to statements in the warehouse's query language, executes these, and formats the results returned into a common, pre-defined representation; iii) an evaluation algorithm computes each rule in ECLECTIC against these results and combines the rule matches according to the Boolean operators to return either a list of patients or patient counts. Applying the approach to a new warehouse requires only a new adaptor for step (ii). A detailed description of the system architecture is included in a prior publication [12].

As each warehouse uses different terminologies, mappings are required between the terminology used in the ECLECTIC query and those used in each warehouse. Code mappings can be looked up automatically from a reference terminology, however we found in some EHR4CR sites that use a specific local terminology, some degree of manual site-specific coding could not be avoided.

4. Results

ECLECTIC has been used in EHR4CR as the language to express ECs, and execute them as queries at clinical sites. A drag-and-drop user interface is employed to compose ECs that is then rendered into ECLECTIC. The EHR4CR platform [13] has been implemented, deployed at 11 clinical sites providing secondary care data, and used by 8 pharma companies to run feasibility studies. EHR4CR uses two different data warehouse schemata to store clinical data. The project developed its own database schema based on the HL7 RIM so that patient data from diverse clinical sites could be loaded into databases of this type. In addition, three sites already had established data warehouses using the i2b2 schema [14]. The ECLECTIC was shown to work with both schemata.

Warehouses vary in content and size from 200 to over 200k patients. Different sets of coding systems are used at different sites and semantic mappings have been implemented to make automated translation possible. The approach is not restricted to the two schemata described above and could be extended to other architectures [15, 16].

A survey of 13 users of the platform revealed that they used ECLECTIC to check a query composed by the GUI was correct, to communicate a set of ECs to a colleague and also to recreate a query on the GUI from the ECLECTIC provided by a colleague. 92% agreed that ECLECTIC was useful as a means of communicating ECs. 77% agreed that it was easy to understand. 85% agreed that it was precise and unambiguous.

```
1 gender() in {[SNOMED Clinical Terms:248153007,"Male"]} and
2 first diagnosis([ICD-10:E11,"Non-insulin-dependent diabetes mellitus"]) and
3 first diagnosis([ICD-10:I50,"Heart failure"]) at least 5 year after rule(2) and
4 not last result([LOINC:4548-4,"Hemoglobin A1c/Hemoglobin.total:Mass Fraction:Point in time:Whole
   blood:Quantitative"]) in range(<=7.0)unit([ucum:%,"percent"])
   at most 3 month before now
```

Figure 1 ECLECTIC example

5. Discussion

Although systems for selecting or counting eligible patients exist, such as i2b2/SHRINE [14] or ePCRN [15], they assume the use of a specific model and/or homogeneous data sources. They also remain limited in respect of temporal constraint evaluations, which are essential for significant number of ECs. A domain-specific query language can overcome these problems by being independent of the particular representation of the data and by supporting those features that may not be present in standard query languages such as temporal semantics.

We would argue that such an approach can be used in other situations where there is a need to allow users to express their queries and to query data from heterogeneous sources for some well-defined purpose while preserving semantic-consistency. Indeed, if there were no commonalities between the clinical data held at clinical sites, then there would be no role for standards such as HL7 RIM and CCR. It is clearly not possible to define a general query language where a single query can access arbitrary database schemata in a semantically consistent way. But where the nature of the queries is well understood in advance and there is a well-defined domain, a domain query language is a viable approach.

6. Conclusions

The query language, ECLECTIC, is powerful enough to express many of the ECs that can be evaluated automatically by the facts held in clinical data warehouses. In particular, it expresses temporal semantics that cannot be achieved practically by standard query languages alone. The simple syntax, CCR-based concepts and use of display names means that ECLECTIC is comprehensible by non-technical users. It has been shown to evaluate patient counts on data warehouses using different structures and coding schemes.

Acknowledgments

This research is 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.This work was part funded by the IMI-funded collaborative project EHR4CR (Electronic Healthcare Records for Clinical Research), Grant agreement no.: 115189. We would like to acknowledge the work on designing and implementing the EHR4CR platform performed by many members of the project.

References

1. Ross J, Tu S, Carini, S, Sim I. Analysis of Eligibility Criteria Complexity in Clinical Trials. AMIA Summits Transl Sci Proc 2010. pp 46 –50.
2. EHR4CR – Electronic Healthcare Records for Clinical Research website. http://www.ehr4cr.eu, last accessed 27/11/2013.
3. Tu S, Peleg M, Carini, S, Bobak M, Ross J, et al. A practical method for transforming free-text eligibility criteria into computable criteria. J Biomed Inform 2011; 44 (2): 239 –250.
4. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review, J Biomed Inform 2010; 43 (3): 451– 467.
5. Wang SJ, Ohno-Machado L, Mar P, Boxwala AA, Greenes RA. Enhancing Arden Syntax for Clinical Trial Eligibility Criteria. Proc AMIA Symp 1999. p 1188.
6. Sordo M, Boxwala A, Ogunyemi O, Greenes R. Description and status update on GELLO: a proposed standardized object-oriented expression language for clinical decision support. Stud Health Technol Inform 2004; 107: 164 –168.
7. Tu S, Peleg M, Carini, Rubin D, Sim I. Ergo – A Template-based Expression Language for Encoding Eligibility Criteria. http://ebookbrowse.com/ergo-technical-documentation-pdf-d47453206 (last accessed 30/4/2013)
8. Benson T. Principles of Health Interoperability HL7 and SNOMED: Chapter 7. Springer; 2009.
9. Standard Specification for Continuity of Care Record (CCR), ASTM E2369 – 05e2, West Conshohocken, PA, USA; 2010.
10. National Quality Forum, Quality Data Model, December 2012. www.qualityforum.org/QualityDataModel.aspx (last accessed 29/11/2013).
11. Nigrin DJ, Kohane IS. Temporal Expressiveness in Querying a Time-stamp-based Clinical Database. J Am Med Inform Assoc 2000; 7 (2): 152 –163.
12. Bache R, Miles S, Taweel A. An Adaptable Architecture for Patient Cohort Identification from Diverse Data Sources. J Am Med Inform Assoc 2013 Sep 24. doi: 10.1136/amiajnl-2013-001858.
13. Chen Y, Bache R, Miles S, Cuggia M, Soto-Rey I, Taweel, A. A SOA-based Platform for Automating Clinical Trial Feasibility Study. In: Proceedings of the IADIS International Conference E-health 2013. IADIS Press; 2013. pp 87–94.
14. i2b2 – Informatics for Integrating Biology and the Bedside. National Centre for Biomedical Computing. https://www.i2b2.org. Accessed 1/6/2013.
15. Ethier JF, Dameron O, Curcin V, McGilchirst M, Verheij R, Arvanitis T, et al. A unified structural/terminological interoperability framework based on LexEVS: application to TRANSFoRm. J Am Med Inform Assoc. 2013; 20 (5): 986 – 994.
16. Delaney BC, Peterson KA, Speedie S, Taweel A, Arvanitis TN, Hobbs FDR. Envisioning a Learning Health Care System: The Electronic Primary Care Research Network. A Case Study. Annals of Family Medicine 10 (1): 54 –59.

