Ontological modelling and FHIR Search based representation of basic eligibility criteria

Ontologische Modellierung und FHIR Search basierte Repräsentation grundlegender Ein- und Ausschlusskriterien

Abstract

Planning clinical studies to check medical hypotheses requires the specification of eligibility criteria in order to identify potential study participants. Electronically available patient data allows to support the recruitment of patients for studies. The Smart Medical Information Technology for Healthcare (SMITH) consortium aims to establish data integration centres to enable the innovative use of available healthcare data for research and treatment optimization. The data from the electronic health record of patients in the participating hospitals is integrated into a Health Data Storage based on the Fast Healthcare Interoperability Resources standard (FHIR), developed by HL7. In SMITH, FHIR Search is used to query the integrated data. An investigation has shown the advantages and disadvantages of using FHIR Search for specifying eligibility criteria. This paper presents an approach for modelling eligibility criteria as well as for generating and executing FHIR Search queries. Our solution is based on the Phenotype Manager, a general ontological phenotyping framework to model and calculate phenotypes using the Core Ontology of Phenotypes.

Keywords: eligibility determination, biomedical ontologies, electronic health records, health information interoperability, information storage, information retrieval, Health Level Seven

Zusammenfassung

Die Planung klinischer Studien zur Überprüfung medizinischer Hypothesen erfordert die Spezifikation von Ein- und Ausschlusskriterien, um potenzielle Studienteilnehmer zu identifizieren. Elektronisch verfügbare Patientendaten ermöglichen es, die Rekrutierung von Patienten für Studien zu unterstützen. Das Konsortium Smart Medical Information Technology for Healthcare (SMITH) hat sich zum Ziel gesetzt, Datenintegrationszentren zu etablieren, um die innovative Nutzung verfügbarer Gesundheitsdaten für Forschung und Behandlungsoptimierung zu ermöglichen. Die Daten aus elektronischen Gesundheitsakten von Patienten in den teilnehmenden Krankenhäusern werden in einem Health Data Storage integriert, der auf dem von HL7 entwickelten Fast Healthcare Interoperability Resources Standard (FHIR) basiert. In SMITH wird FHIR Search verwendet, um die integrierten Daten abzufragen. Eine Untersuchung hat die Vor- und Nachteile der Verwendung von FHIR Search zur Spezifikation von Ein- und Ausschlusskriterien aufgezeigt. Dieser Artikel präsentiert einen Ansatz zur Modellierung von Ein- und Ausschlusskriterien sowie zur Generierung und Ausführung von FHIR Search Queries. Unsere Lösung basiert auf dem Phenotype Manager, einem allgemeinen ontologischen Phänotypisierungs-Framework zur Modellierung und Berechnung von Phänotypen unter Verwendung der Core Ontology of Phenotypes.

Alexandr Uciteli[1,2]
Christoph Beger[1,3]
Jonas Wagner[1,2,4]
Toralf Kirsten[2,4,5]
Frank A. Meineke[1,2]
Sebastian Stäubert[1,2]
Matthias Löbe[1,2]
Heinrich Herre[1,2]

1 Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Germany

2 SMITH Consortium of the German Medical Informatics Initiative

3 Growth Network CrescNet, University of Leipzig, Germany

4 LIFE Research Centre for Civilization Diseases, University of Leipzig, Germany

5 Faculty of Applied Computer and Biological Sciences, University of Applied Sciences Mittweida, Germany

Introduction

Aim

Planning clinical studies to check medical hypotheses requires the specification of eligibility criteria in order to identify potential study participants. Electronically available patient data allows to support the recruitment of patients for studies [1], [2]. A major challenge is the difficulty to match subject eligibility criteria to query capabilities of the electronic health record repositories [3]. The applicability of queries can be improved by different strategies, such as separating combined criteria into simpler ones or using controlled terminologies, libraries of shared queries and interactive data entry form applications [3].

The German Medical Informatics Initiative (MII) [4], [5] aims to make clinical data available for research. In order to overcome the heterogeneity of the data, a core data set [6] is being developed. On the basis of the core data set, the data will be structured in the same way in all participating institutions. Before a scientist can use the data in analysis projects, a possibility to determine the number of potentially eligible patients (e.g., by a query in a portal) would be desirable. The Federal Ministry of Education and Research funded multiple consortia involving most of the German university hospitals. Smart Medical Information Technology for Healthcare (SMITH) is one of these consortia [7]. The objective of the SMITH consortium is to establish data integration centres to provide researchers with available healthcare data and knowledge and to enable the innovative use of data for research and treatment optimization. The data from electronic health records (EHR) of patients in the participating hospitals is integrated into a so-called Health Data Storage (HDS) in a standardized manner using the HL7 standard Fast Healthcare Interoperability Resources (FHIR) [8]. A natural language processing (NLP) pipeline is developed to extract and transform relevant data from unstructured EHR documents into structured form. The HDS can be accessed from outside via a FHIR server. In SMITH, FHIR Search [9] is used to query the integrated data. Once the patient data is available in a structured and standardized form, scientists can use a data portal to execute their eligibility criteria (feasibility) queries on the HDS and obtain the number of relevant cases according to the scientific hypothesis.

The capabilities of FHIR Search for specifying eligibility criteria were investigated by Gulden et al. [10] based on 25 studies from ClinicalTrials.gov. The investigation has shown that FHIR Search is helpful in supporting pre-screening efforts, but it has also some limitations.

The methodical use case 'phenotyping pipeline' (PheP) [11] of the SMITH project provides a systematic approach to develop, evaluate and execute validated phenotype algorithms for classifying patients based on routine EHR data. In the context of PheP, we develop the software Phenotype Manager (PhenoMan), a general ontological phenotyping framework to model and calculate phenotypes based on patient data from HDS using the Core Ontology of Phenotypes (COP) [12], [13]. We consider eligibility criteria as one aspect in this framework, and a set of criteria as a specification of a phenotype class.

In this article, we present an ontological approach for modelling basic inclusion and exclusion criteria as well as for generating and executing FHIR Search queries based on the PhenoMan framework. We analyse limitations of FHIR Search and propose possible workarounds. Our work serves as conceptual preliminary work for the development of the SMITH data portal.

Problem statement

In this section, we outline the limitations of FHIR Search described by Gulden et al. [10] (li1-li4a) as well as our observations (li4b-li6) resulting from a detailed analysis of the FHIR Search specification.

- (li1) The FHIR Search specification [9] does not support 'age' as search parameter. However, extensions can be developed that overcome this problem.
- (li2) Complex queries with inter-data dependencies and necessary computations (like 'children whose BMI is less than the 95th percentile for their age, with children between 2 and 5 years old' [10]) cannot be expressed. The required values have to be calculated first, before the resulting query can be specified and executed.
- (li3) Representing both inclusion and exclusion criteria in a single query is difficult. Instead, intersecting patients between the inclusion and exclusion criteria need to be explicitly removed from the final result set.
- (li4a) Complex temporal constraints are difficult to represent (e.g., searching for patients who had a certain procedure in a given period of time). Additional exclusion criteria queries may be required to exclude irrelevant periods of time.
- (li4b) The limitation (li4a) is related to the usage of the _has parameter (e.g., patient that _has some characteristics or observations), which can cause some problems. According to the FHIR Search specification [9], each _has parameter is processed independently of other _has parameters. It is therefore, for example, impossible to express that a patient had a certain observation in a specific period of time. If we use, e.g., the following query:

  Patient?_has:Observation:patient:code=http://snomed.info/sct|27113001&_has:Observation:patient:date=gt2019-09-30

  we do not search for patients that have a weight observation (http://snomed.info/sct 27113001) after the 30th of September 2019, but for patients that have a weight observation at an arbitrary date AND an arbitrary observation after the 30th of September 2019 (e.g., a weight observation in September and a height observation in November).
- (li5) Value range restrictions can only be specified if the measurement units of the corresponding charac-

teristics are known. For instance, if we do not know which unit was used for height observations (e.g., meters or centimeters), we cannot define the restriction. The conversion of units does not take place automatically.

- (li6) Multiple value range restrictions of a characteristic (e.g., age between 20–40 or 60–80 years) can not be expressed using one query. One query for each value range has to be specified.

Methods

Basic inclusion/exclusion criteria

The integration of mathematical or statistical calculations in the query syntax is out of the scope of this paper (li2). We assume that all required values are calculated and recorded in the electronic health record (EHR).

We consider only the following kinds of eligibility criteria:

Inclusion criteria:

- (ic1) a given patient characteristic must occur in the EHR (e.g., a diabetes diagnosis must be present)
- (ic2) a given patient characteristic must occur in the EHR and must be observed in the specific period of time (e.g., a diabetes diagnosis must be observed in the last month)
- (ic3) a given patient characteristic must occur in the EHR and must have a value that lies within a defined value range (e.g., the weight value must be between 70 and 80 kg)

Exclusion criteria:

- (ec1) a given patient characteristic must not occur in the EHR (e.g., a diabetes diagnosis must not be present)
- (ec2) a given patient characteristic may occur in the EHR but must not be observed in the specific period of time (e.g., a diabetes diagnosis must not be observed in the last month)
- (ec3) a given patient characteristic may occur in the EHR but must not have a value that lies within a defined value range (e.g., the weight value must not be between 70 and 80 kg)

Our approach focuses on the development of an algorithm that supports the specification of such eligibility criteria, generation and execution of required queries as well as the collection of relevant information about patients meeting all criteria.

FHIR Search

The FHIR Search Framework [9] is part of the HL7 FHIR standard and provides a range of operations and parameters (series of name=value pairs) to search for existing FHIR resources in the underlying repository. In the simplest case, a search is executed by performing a GET operation in the RESTful framework:

```
GET [base-url]/[resource-type]?name=value&...
{&_format=[mime-type]}}
```
e.g., *GET [base-url]/Patient?gender=male*

Some parameters, such as _content, _id, _lastUpdated, etc. can be applied for all resource types, while other parameters hold only for specific resource types (e.g., "value-quantity" for Observation).

For numeric parameter types (number, date or quantity), a value range can be defined using a prefix to the parameter value (e.g., gt=greater than, le=less or equal).

The '&' (AND) operator between single search criteria is used to search for the intersection of resources that match all criteria specified by each individual search parameter (e.g., *Patient?gender=male&birthdate= gt1970*). To search for resources with one of the specified parameter values (OR), the values must be separated by a comma (e.g., *Observation?code=http://loinc.org| 3141-9,http://snomed.info/sct|27113001*, i.e., weight code from LOINC or SNOMED).

The following query contains AND combinations of single criteria (code AND value-quantity AND date) as well as OR linking of code values and can be used to search for weight observations that are not older than October 2018 and where the weight is greater than 75 kg:

Observation?code=http://loinc.org|3141-9, http://snomed.info/sct|27113001&value-quantity=gt75||kg&date=gt2018-09

Composite search parameters allow a search based on a pair of values. The both parameters 'component-code' and 'value-quantity', for example, can be combined to one parameter 'component-code-value-quantity'. The query for the observation of systolic blood pressure ≥130 mmHg in the last year is as follows:

Observation?component-code-value-quantity= http://snomed.info/sct|271649006$ ge130|http://unitsofmeasure.org|mm[Hg]&date= ge2019-04-24

The _has parameter provides limited support for reverse chaining. It enables to select resources based on the properties of resources that refer to them. For instance, the following query allows to search for patients that have a weight observation:

Patient?_has:Observation:patient:code=http:// snomed.info/sct|27113001

Each _has parameter is processed independently of other _has parameters (li4b).

Results

Concept

In this section, we address the outlined limitations of FHIR Search and propose a strategy for specifying and executing eligibility criteria queries.

We assume that a set of eligibility criteria is given and we consider the criteria as AND-linked, i.e., we search for patients that meet all the criteria. To prevent complications, we propose to avoid using the _has parameter (li4).

The _has parameter enables to search for patients that have certain characteristics without the possibility to restrict some of them. In this way, only the (ic1) criteria can be processed. To manage the (ic2) or (ic3) criteria, the result set has to be filtered afterwards based on the defined restrictions (period of time or value range). This functionality must then be implemented in the client software. Furthermore, the exclusion criteria cannot be expressed using the _has parameter, because the _has parameter cannot be negated.

Our approach consists of the following steps:

1. Specify and execute a query for patients with gender and age defined as inclusion criteria. To overcome the (li1), server-side extensions can be used/developed that support the 'age' parameter. Alternatively, the transformation of 'age' to 'birth date' can be implemented on the client side. The resulting patients constitute the initial patient set (PS). We only need the patient ID, so that not much data has to be transmitted. For example:
   Patient?gender=male&birthdate=gt1970
   Note: The patient query can be extended by other inclusion criteria characteristics (without value range restrictions) using the _has parameter, which may lead to a reduction of the initial patient set. However, our investigations have shown that it has no advantages. The subsequent reduction of the patient set is more efficient than the use of the time-consuming _has parameter.
2. Specify and execute a query for each inclusion criterion including specific restrictions. For example (weight >75 kg, date > September 2018):
   Observation?code=http://loinc.org|3141-9, http://snomed.info/sct|27113001&value-quantity= gt75||kg&date=gt2018-09
   If a characteristic has multiple restrictions (e.g., weight >75 kg or weight <45 kg), build multiple queries.
   Note: If the measurement units are unknown (li5) or there are multiple value range restrictions (li6), the value range checks should be performed by the client software.
   If no inclusion criteria were defined, continue with step 5.
3. After executing each inclusion criterion query, reduce the PS to patients with a returned characteristic (set-theoretical intersection) (li3). This means that all patients that do not meet the criterion (no data returned) are removed from the PS.
4. If the PS is empty, terminate the execution with the output 'no patients meet the inclusion criteria'.
5. If the PS is not empty, specify and execute a query for each exclusion criterion including specific restrictions (similarly to 2.).
6. Remove patients with resulting resources (5.) from the PS (set-theoretical difference) (li3).
7. Output the PS.

The main goal of our solution is to minimize the number of required queries. In contrast to a patient-centred approach, we apply a characteristic-centred one, i.e., instead of querying all relevant characteristics for each patient, we search for the characteristics first and then associate them with patients. Additionally, the possibility to check value range restrictions in the client software can potentially reduce the number of queries. Thus, the maximum number of queries is limited to the number of the eligibility criteria plus the one query for patients (with defined age and gender).

Implementation

In the SMITH project, we develop the PhenoMan, an ontology-based software for modelling, classification and calculation of phenotypes [13]. Phenotypes are specified and saved in OWL ontologies using PhenoMan. The detailed structure of the ontology and our definition of the phenotype notion is presented in [12], [13].

For specifying eligibility criteria, only the 'atomic' or single phenotypes [13] are relevant. Single phenotypes are single properties (e.g., age, weight or height) of an organism or of one of its subsystems. To ontologically model single phenotypes, we use non-restricted single phenotype (NSiP) and restricted single phenotype (RSiP) classes. For example, the phenotype class 'age' is instantiated by the ages of all living beings (non-restricted), whereas the phenotype class 'young age' is instantiated by the ages of the young ones, e.g., if the age is below 30 years (restricted). RSiP classes are subclasses of NSiP classes, e.g., the class 'young age' is a subclass of the class 'age'. Phenotype classes possess various common attributes (e.g., labels, descriptions and codes of external terminologies such as LOINC or SNOMED CT). Additionally, the NSiP classes define the data type, a unit of measure, a validity period, a corresponding resource type (e.g., FHIR Observation, Condition or Procedure) and an optional aggregate function. The PhenoMan supports 4 primitive data types xsd:decimal, xsd:string, xsd:boolean and xsd:date. Complex data types (e.g., FHIR code or quantity) are mapped to the primitive data types (e.g., code to xsd:string and quantity to xsd:decimal with additional unit attribute). The RSiP classes specify a value range restriction of a corresponding data type. We distinguish between enumerated and limited value ranges. An enumerated value range is used to express a predefined set of permitted values, such as the value set 'administrative gender' [14] defined by the FHIR project. The limited value ranges define an upper and a lower limit (including or excluding, e.g., weight >60 and ≤80 kg).

During specification, the NSiP classes can be marked as inclusion or exclusion criteria. The PhenoMan uses this information to execute the algorithm specified in the section Concept. If a NSiP class has no subclasses (i.e., no RSiP classes) and defines no validity period, it can only be processed as (ic1) or (ec1), depending on whether the class was specified as inclusion or exclusion criterion. In this case, it will only be checked whether the appropriate data is present. If a NSiP class additionally specifies a validity period, it will be handled as (ic2) or
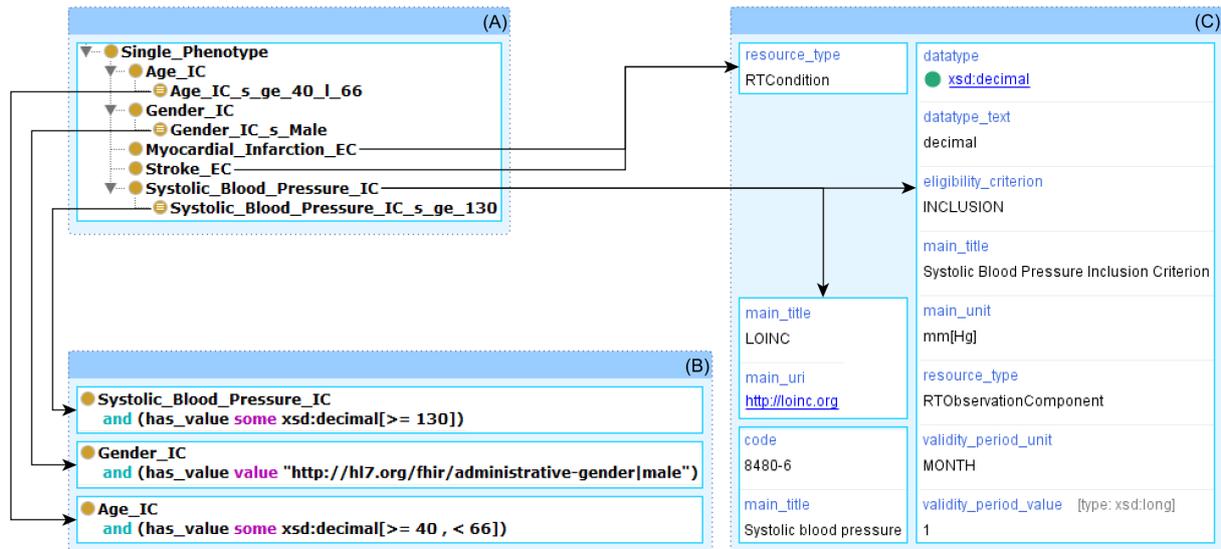
Figure 1: Specification of eligibility criteria in Protégé [36] (excerpt)
(A) Ontology tree: subclasses of Single_Phenotype are NSiP classes (e.g., Age_IC); their subclasses are RSiP classes (e.g., Age_IC_s_ge_40_l_66). (B) Specification of value range restrictions as anonymous equivalent classes of RSiP classes. The gender is restricted using the code 'male' from the FHIR value set 'administrative gender'. (C) Example annotations of NSiP classes (e.g., code, data type, title, resource type, measurement unit, validity period).

(ec2). Finally, if a NSiP class has a restriction, i.e., a RSiP class, it will be managed as (ic3) or (ec3).

The PhenoMan implements the approach described above. It transforms the age into the birth date, generates and executes required FHIR Search queries and calculates the resulting patient set meeting all specified criteria. During generation of queries, the defined codes, value range restrictions, units of measure, validity periods and resource types are taken into account. For FHIR Observations, a distinction is made between single (e.g., weight observation) and composite observations (e.g., blood pressure observation consisting of a systolic and a diastolic blood pressure value). Additionally, the PhenoMan supports the conversion of measurement units, specified using the Unified Code for Units of Measure (UCUM) [15]. We outline the functionality of PhenoMan using a simple example. For a blood pressure study, the following eligibility criteria were defined:

Inclusion criteria:

- male subjects
- age between 40 and 65 years (i.e., age in years ≥40 and <66)
- at least one observation of systolic blood pressure ≥130 mmHg in the last month

Exclusion criteria:

- myocardial infarction
- stroke

The eligibility criteria are modelled using the Phenotype Editor of the PhenoMan [13] and are saved in the ontology. The resulting ontology [16] (excerpt) is represented in Figure 1.

As a first step, the PhenoMan transforms the age range into the corresponding birth date range (li1). Then, the patient query is generated (in consideration of birth date and gender) and sent to the FHIR server:

Patient?birthdate=gt1954-04-24&birthdate=le1980-04-24&gender=male

The obtained patients constitute the initial patient set (PS). Next, the PhenoMan generates and executes the blood pressure query with the value range restriction and validity period:

Observation?component-code-value-quantity= http://loinc.org|8480-6$ ge130|http://unitsofmeasure.org|mm[Hg]&date= ge2020-03-24T14:03:42

Alternatively, the PhenoMan can check the value range restrictions (li5, li6).

The PS is reduced to patients with a returned valid blood pressure observation (intersection). Finally, the PhenoMan generates and executes both exclusion criteria queries:

Condition?code=http://snomed.info/sct|22298006 Condition?code=http://snomed.info/sct|230690007

and removes (set-theoretical difference) patients with returned conditions (myocardial infarction or stroke) from the PS. Since in this case, no data restrictions are required and only condition codes are relevant, both queries can be combined to one OR query:

Condition?code=http://snomed.info/sct|22298006, http://snomed.info/sct|230690007

Evaluation

The FHIR Search specification and the functionality of the HAPI FHIR [17] server used for the evaluation are beyond the scope of this paper. We assume that the HAPI FHIR server as well as the underlying data repository correctly implement the FHIR specification. We consider FHIR as an established technology (at least in our project)

and it is not our goal to compare the general functionality or performance of FHIR Search with, e.g., SQL. Therefore, the variability of the test queries and the real patient data have no influence on our approach. Few queries of each supported type (patient query, ic/ec 1–3 queries) as well as generated patient data are sufficient for the evaluation. The usability of the Phenotype Editor [13] to specify phenotypes (eligibility criteria) will be investigated in further papers.

Our evaluation focussed on

1. generation of correct queries from the ontological specification of corresponding eligibility criteria and
2. accurate execution of the algorithm described in the Concept section, i.e., correct determining patients satisfying all criteria.

For the evaluation, we used realistic patient data created by the synthetic patient generator Synthea [18]. The open-source software models the medical history of synthetic patients and provides high-quality realistic patient data and associated health records covering every aspect of healthcare. We generated 66,018 virtual patients with corresponding health records as FHIR bundles (JSON). The FHIR bundles were imported in the HAPI FHIR server and saved in the integrated PostgreSQL database. The underlying database schema is described in [19]. The generated data could then be queried using the FHIR Search interface of the HAPI FHIR server as well as directly on the PostgreSQL database via SQL.

The example from the Implementation section was chosen for the evaluation, because it contains all relevant query types supported by PhenoMan, such as:

- patient query:
  Patient?birthdate=gt1954-04-24&birthdate=le1980-04-24&gender=male
- ic1/ec1 query:
  Condition?code=http://snomed.info/sct|22298006, http://snomed.info/sct|230690007
- ic2/ic3/ec2/ec3 query:
  Observation?component-code-value-quantity= http://loinc.org|8480-6$ ge130|http://unitsofmeasure.org|mm[Hg]&date= ge2020-03-24T14:03:42

To compare our approach, we developed SQL queries for the PostgreSQL database as a gold standard (Figure 2). Table 1 summarises the evaluation results. The matching subjects of each query pair were identical. Although the performance of our approach or of FHIR Search is not a critical issue in our use case (e.g., the queries could run overnight), we measured the execution time of the queries.

The evaluation results and the example queries are available at [20], the generated health records at [21].

Discussion

The topic 'FHIR' and especially 'FHIR Search' is relatively new. In April 2020, the query 'FHIR Search' delivered on Google Scholar 26 publications and the query '"FHIR Search" "Eligibility Criteria"' even only one paper, namely [10] that we have already mentioned in the chapter Methods. Most publications relevant in the context of this work deal with i2b2 [22], SMART-on-FHIR [23] or a combination of i2b2 and SMART-on-FHIR [24], [25].

SMART-on-FHIR offers a FHIR-based specification and a reference implementation of an interoperable apps platform for electronic health records [23]. It defines a way for health apps to connect EHR systems with appropriate security guarantees and extends the FHIR specification by authorization, authentication and UI integration components. An interface (SMART-on-FHIR cell) was developed to serve patient data from i2b2 repositories in FHIR format [25]. The cell delivers FHIR resources on a per-patient basis and supports the SMART specification. A comparable implementation of a FHIR layer (FHIR server) over i2b2 is presented by Boussadi et al. [22]. It is based on a locally developed FHIR profile and HAPI FHIR API and is not limited to serve data on a per-patient basis. The most important difference between these approaches and our solution is that we integrate the data directly in a FHIR conform data storage and use the native query language (FHIR Search) without the need to transform the data.

Similar to our solution, Paris et al. [24] presented an i2b2 extension to search remotely in FHIR repositories (using SMART-on-FHIR APIs). This enables the federation of queries, security, terminology mapping, and also bridges the gap between i2b2 and modern big-data technologies. The existing, traditional i2b2 query search module was extended to meet the SMART-on-FHIR API specifications and the FHIR Search specifications. In contrast to this approach, we model the eligibility criteria ontologically and use the ontology for generating FHIR Search queries. An interesting solution to specify patient selection criteria using ART-DECOR [26] was presented by Ott el al. [27]. Existing ART-DECOR template associations have been extended by the possibility to specify the required conditions. The defined criteria are translated to XPath expressions and can be applied to CDA documents to identify eligible patients. Similarly to [27], in the project SMITH, metadata elements (including single data elements, data element groups, value sets, referenced terminologies, etc.) are specified using ART-DECOR. Our approach is to access the specified data elements via ART-DECOR API and to integrate them into the ontology as non-restricted single phenotype classes [13]. Then, the complex phenotypes are modelled based on the integrated data elements using the Phenotype Editor and are saved in the ontology [13]. The ontology is used by PhenoMan to generate and execute FHIR Search queries.

A number of groups are developing knowledge representation formalisms for eligibility criteria [28].

```
WITH patients AS (
            -- inclusion: male subjects
            -- inclusion: age >= 40 and <= 65 years
            SELECT DISTINCT res_id
            FROM hfj_spidx_token t
                    JOIN hfj_spidx_date d USING (res_id, res_type)
            WHERE res_type = 'Patient'
                    AND t.sp_name = 'gender'
                    AND sp_value = 'male'
                    AND d.sp_name = 'birthdate'
                    AND extract(year FROM age(sp_value_high)) BETWEEN 40 AND 65
), observations AS (
            -- inclusion: observation of systolic blood pressure >= 130 mmHg in the last month
            SELECT DISTINCT res_id
            FROM hfj_spidx_date d
                JOIN (
                        SELECT res_id, sp_name, sp_value
                        FROM hfj_spidx_token
                        WHERE sp_name = 'component-code'
                ) t USING (res_id)
                JOIN (
                        SELECT res_id, sp_name, sp_value
                        FROM hfj_spidx_quantity
                        WHERE sp_name = 'component-value-quantity'
                ) q USING (res_id)
            WHERE res_type = 'Observation'
                    AND d.sp_name = 'date'
                    AND d.sp_value_high >= CURRENT_DATE - interval '1 months'
                    AND t.sp_value = '8480-6'
                    AND q.sp_value >= 130
), exclusions AS (
            -- exclusion: myocardial infarction or stroke
            SELECT DISTINCT res_id
            FROM hfj_spidx_token t
            WHERE res_type = 'Condition'
                    AND t.sp_value IN ('22298006', '230690007')
)
SELECT DISTINCT res_id
FROM (
            SELECT p.res_id
            FROM observations o
                    JOIN hfj_res_link ON (o.res_id = src_resource_id)
                    JOIN patients p ON (p.res_id = target_resource_id)
            EXCEPT
            SELECT target_resource_id
            FROM exclusions
                    JOIN hfj_res_link ON (res_id = src_resource_id)
            WHERE target_resource_type = 'Patient'
) q;
```

Figure 2: Example gold standard SQL query

Tu et al. [29] used the Eligibility Rule Grammar and Ontology (ERGO) to annotate free-text criteria and to transform them in a computable form.

Chondrogiannis et al. [30] proposed a CDISC-compliant schema for organizing criteria along with a patient-centric model for their formal expression, properly linked with international classifications and codification. The Eligibility Criteria Ontology (EC-O) has been developed to cover a wide range of parameters allowing the specification of eligibility criteria in clinical trials. EC-O contains seven main eligibility criteria categories (Demographics, Diagnosis, Laboratory examination, etc.) as well as three different types of their relations with a person. The eligibility criteria are formally represented in XML or SPARQL. The XML schema as well as the structure of the SPARQL queries are based on the EC-O. The XML-based expressions are manually specified using a GUI tool, whereas the SPARQL-based expressions can be generated automatically.

Table 1: Evaluation results

Query No.	Inclusion	Exclusion	Hits [Count]	PhenoMan [s]	SQL [s]
1	Male, age ≥40 and <66 years, at least one observation of systolic blood pressure ≥130 mmHg in the last month	Myocardial infarction, stroke	113	309	51
2	Female, age ≥40 and <60 years, BMI ≥34 and <70		119	302	50
3	Male, age ≥40 and <60 years, diabetes mellitus type 2, BMI ≥35 and <70	Smokes tobacco daily	6	429	157
4	Female, age ≥40 and <60 years, diabetes mellitus type 2, Haemoglobin A1c/Haemoglobin total in Blood ≥6%		684	361	153
5	Male, age ≥40 and <60 years, diabetes mellitus type 2, BMI ≥26 and <35 in the last month	Anaemia, myocardial infarction, atrial fibrillation, stroke	8	526	257
6	Female, age ≥18 and <40 years, asthma	Smokes tobacco daily	10	98	40
7	Male, age ≥18 and <40 years, allergy to animal dander		18	15	8
8	Female, age ≥18 and <60 years, acute bronchitis, allergy to animal dander	Smokes tobacco daily	11	174	14
9	Male, age ≥18 and <60 years, acute bronchitis, allergy to grass pollen		4	265	31
10	Age ≥18 and <40 years, allergy to grass pollen	Acute bronchitis, asthma	6	327	28

An approach of making eligibility criteria computable by using the ontology-based data access framework Ontop [31] is presented in [32]. The Ontology for Computable Eligibility Criteria was constructed to represent the eligibility criteria commonly used in Hepatitis C trials. Templates of SPARQL queries were designed based on the eligibility criteria use cases. The SPARQL queries are utilized to query relational databases as virtual RDF graphs.

We consider modelling and execution of eligibility criteria as one aspect in a general ontological phenotyping framework, and a set of criteria as a specification of a phenotype class [13]. The successful determination and analysis of phenotypes plays a key role in the diagnostic process, the evaluation of risk factors and the recruitment of participants for clinical and epidemiological studies. We believe that a broad ontologically oriented view is useful to achieve a semantically correct representation of phenotypes and to support the acquisition, representation and retrieval of complex medical data.

Our approach focuses on the use of FHIR Search to support a simple specification of queries by scientists using a graphical tool. Nonetheless, we investigate the future application scenarios of other more expressive alternatives. FHIRPath [33] is a path-based navigation and extraction language similar to XPath. Operations are expressed in the form of logical operations on a hierarchical data model and support traversing, selecting and filtering data. FHIRPath is specified independently of FHIR, but supports the full functionality of FHIR Search. Another standard for complex queries on FHIR data is the Clinical Quality Language (CQL) [34]. CQL defines a language for querying clinical knowledge that can be used primarily for Clinical Decision Support (CDS) and Clinical Quality Measurement (CQM). The Clinical Reasoning Module of FHIR [35] specifies the use of the Clinical Quality Language within FHIR and also the method to represent knowledge artefacts. CQL is a superset of FHIRPath, i.e., any valid FHIRPath expression is also a valid CQL expression. Path navigation in document-like data structures can thus also be expressed.

Future work also includes the implementation of more complex types of eligibility criteria as well as complex Boolean connections between individual criteria, e.g., using the _filter parameter of FHIR Search (the modelling in the ontology is already functional).

Conclusion

We developed an ontology-based approach for specification of eligibility criteria as well as generation and execution of required FHIR Search queries to retrieve relevant information about patients meeting all criteria. Our method is embedded in a general phenotyping framework to model and calculate phenotypes based on patient data (e.g., from electronic health records). The framework is developed and applied in the context of a methodological use case 'phenotyping pipeline' of the German Medical Informatics Initiative project SMITH.

Notes

Competing interests

The authors declare that they have no competing interests.

Funding

References

1. Dugas M, Lange M, Müller-Tidow C, Kirchhof P, Prokosch HU. Routine data from hospital information systems can support patient recruitment for clinical studies. Clin Trials. 2010 Apr;7(2):183-9. DOI: 10.1177/1740774510363013

2. Köpcke F, Kraus S, Scholler A, Nau C, Schüttler J, Prokosch HU, Ganslandt T. 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 Inform. 2013 Mar;82(3):185-92. DOI: 10.1016/j.ijmedinf.2012.11.008

3. Wang AY, Lancaster WJ, Wyatt MC, Rasmussen LV, Fort DG, Cimino JJ. Classifying Clinical Trial Eligibility Criteria to Facilitate Phased Cohort Identification Using Clinical Data Repositories. AMIA Annu Symp Proc. 2017;2017:1754-63.

4. Gehring S, Eulenfeld R. German Medical Informatics Initiative: Unlocking Data for Research and Health Care. Methods Inf Med. 2018 Jul;57(S 01):e46-e49. DOI: 10.3414/ME18-13-0001

5. Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med. 2018 Jul;57(S 01):e50-e56. DOI: 10.3414/ME18-03-0003

6. Medizininformatik-Initiative. MI-I-Kerndatensatz [Internet]. 2017. Available from: https://www.medizininformatik-initiative.de/de/kerndatensatz

7. Winter A, Stäubert S, Ammon D, Aiche S, Beyan O, Bischoff V, Daumke P, Decker S, Funkat G, Gewehr JE, de Greiff A, Haferkamp S, Hahn U, Henkel A, Kirsten T, Klöss T, Lippert J, Löbe M, Lowitsch V, Maassen O, Maschmann J, Meister S, Mikolajczyk R, Nüchter M, Pletz MW, Rahm E, Riedel M, Saleh K, Schuppert A, Smers S, Stollenwerk A, Uhlig S, Wendt T, Zenker S, Fleig W, Marx G, Scherag A, Löffler M. Smart Medical Information Technology for Healthcare (SMITH). Methods Inf Med. 2018 Jul;57(S 01):e92-e105. DOI: 10.3414/ME18-02-0004

8. HL7. FHIR [Internet]. Available from: https://www.hl7.org/fhir/

9. HL7. FHIR Search [Internet]. Available from: https://www.hl7.org/fhir/search.html

10. Gulden C, Mate S, Prokosch HU, Kraus S. Investigating the Capabilities of FHIR Search for Clinical Trial Phenotyping. Stud Health Technol Inform. 2018;253:3-7. DOI: 10.3233/978-1-61499-896-9-3

11. Meineke FA, Stäubert S, Löbe M, Uciteli A, Löffler M. Design and Concept of the SMITH Phenotyping Pipeline. Stud Health Technol Inform. 2019 Sep;267:164-72. DOI: 10.3233/SHTI190821

12. Herre H, Beger C, Uciteli A. Core Ontology of Phenotypes [Internet]. Leipzig Health Atlas; 2020. Available from: https://health-atlas.de/lha/81GYG30326-8

13. Uciteli A, Beger C, Kirsten T, Meineke FA, Herre H. Ontological Modelling and Reasoning of Phenotypes. In: Barton A, Seppälä S, Porello D, editors. Proceedings of the Joint Ontology Workshops (JOWO) 2019. Episode V: The Styrian Autumn of Ontology; 2019 Sep 23-25; Graz, Austria. (CEUR Workshop Proceedings; 2518). Available from: http://ceur-ws.org/Vol-2518/paper-ODLS11.pdf

14. HL7. Value Set Administrative Gender - FHIR v4.0.1 [Internet]. Available from: https://www.hl7.org/fhir/valueset-administrative-gender.html

15. Unified Code for Units of Measure (UCUM) [Internet]. Available from: https://unitsofmeasure.org

16. Uciteli A. Eligibility Criteria Ontology for an Example Blood Pressure Study [Internet]. Leipzig Health Atlas; 2020. Available from: https://health-atlas.de/lha/8226WG044E-6

17. Smile CDR. HAPI FHIR - The Open Source FHIR API for Java [Internet]. Available from: https://hapifhir.io/

18. MITRE Corporation. Synthea [Internet]. Available from: https://synthetichealth.github.io/synthea/

19. HAPI FHIR Database Schema [Internet]. Available from: https://hapifhir.io/hapi-fhir/docs/server_jpa/schema.html

20. Onto-Med/phenoman-evaluation [Internet]. Available from: https://github.com/Onto-Med/phenoman-evaluation

21. Beger C, Uciteli A. PhenoMan Evaluation with Synthetic FHIR Data [Internet]. Leipzig Health Atlas; 2020. Available from: https://www.health-atlas.de/lha/82UY638KN5-7

22. Boussadi A, Zapletal E. A Fast Healthcare Interoperability Resources (FHIR) layer implemented over i2b2. BMC Med Inform Decis Mak. 2017 Aug;17(1):120. DOI: 10.1186/s12911-017-0513-6

23. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc. 2016 Sep;23(5):899-908. DOI: 10.1093/jamia/ocv189

24. Paris N, Mendis M, Daniel C, Murphy S, Tannier X, Zweigenbaum P. i2b2 implemented over SMART-on-FHIR. AMIA Jt Summits Transl Sci Proc. 2018;2017:369-78.

25. Wagholikar KB, Mandel JC, Klann JG, Wattanasin N, Mendis M, Chute CG, Mandl KD, Murphy SN. SMART-on-FHIR implemented over i2b2. J Am Med Inform Assoc. 2017 Mar;24(2):398-402. DOI: 10.1093/jamia/ocw079

26. ART-DECOR® [Internet]. Available from: https://www.art-decor.org/

27. Ott S, Rinner C, Duftschmid G. Expressing Patient Selection Criteria Based on HL7 V3 Templates Within the Open-Source Tool ART-DECOR. Stud Health Technol Inform. 2019;260:226-33.

28. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. J Biomed Inform. 2010 Jun;43(3):451-67. DOI: 10.1016/j.jbi.2009.12.004

29. Tu SW, Peleg M, Carini S, Bobak M, Ross J, Rubin D, Sim I. A practical method for transforming free-text eligibility criteria into computable criteria. J Biomed Inform. 2011 Apr;44(2):239-50. DOI: 10.1016/j.jbi.2010.09.007

30. Chondrogiannis E, Andronikou V, Tagaris A, Karanastasis E, Varvarigou T, Tsuji M. A novel semantic representation for eligibility criteria in clinical trials. J Biomed Inform. 2017 May;69:10-23. DOI: 10.1016/j.jbi.2017.03.013

31. Calvanese D, Cogrel B, Komla-Ebri S, Kontchakov R, Lanti D, Rezk M, Rodriguez-Muro M, Xiao G. Ontop: Answering SPARQL queries over relational databases. Semantic Web. 2017 Jan 1;8(3):471-87. DOI: 10.3233/SW-160217

32. Zhang H, He Z, He X, Guo Y, Nelson DR, Modave F, Wu Y, Hogan W, Prosperi M, Bian J. Computable Eligibility Criteria through Ontology-driven Data Access: A Case Study of Hepatitis C Virus Trials. AMIA Annu Symp Proc. 2018;2018:1601-10.

33. HL7. FHIRPath - FHIR v4.0.1 [Internet]. Available from: https://www.hl7.org/fhir/fhirpath.html

34. HL7. Clinical Quality Language (CQL) [Internet]. Available from: https://cql.hl7.org/

35. HL7. Clinical Reasoning Module - FHIR v4.0.1 [Internet]. Available from: http://www.hl7.org/fhir/clinicalreasoning-module.html

36. Protégé: A free, open-source ontology editor and framework for building intelligent systems [Internet]. Available from: http://protege.stanford.edu/

Corresponding author:

Alexandr Uciteli
Institut für Medizinische Informatik, Statistik und Epidemiologie, Härtelstraße 16-18, 04107 Leipzig, Germany
auciteli@imise.uni-leipzig.de