Good Evaluation - Better Digital Health
U.H. Hübner et al. (Eds.)

doi:10.3233/SHTI251519

A FHIR Specification to Formalize Cohort Definitions

Britta BERENS[a], Joscha GRÜGER[b,c], Carolin POSCHEN[a] and Konstantin KNORR[a,1]
[a]Department of Computer Science, Trier University of Applied Sciences, Trier, Germany
[b]German Research Center for Artificial Intelligence (DFKI), Branch Trier, Trier, Germany
[c]Business Information Systems II, University of Trier, Trier, Germany
ORCiD ID: Britta Berens https://orcid.org/0000-0002-0324-9640, Joscha Grüger https://orcid.org/0000-0001-7538-1248, Carolin Poschen https://orcid.org/0009-0003-8272-663X, Konstantin Knorr https://orcid.org/0009-0000-2460-5053

Abstract. Retrospective studies play an important role in advancing medical research, yet especially cohort definitions are often provided in unstructured text form or in a non-standardized format. This lack of formalization hinders reproducibility, consistency, and automated reuse. Therefore, we present a framework for the structured and standardized specification of cohort definitions within FHIR resources. Drawing from a systematic review of retrospective studies, we derived six modelling categories for cohort definitions (1) patient demographics, (2) standardized medical terminology, (3) clinical results definition, (4) temporal data representation, (5) temporal relationships and dependencies, and (6) logical combination of criteria. Each category is implemented using native or minimally extended FHIR properties, establishing a one-to-one correspondence between cohort definitions and the original clinical data. This enables both human readability and automated processing, supporting use cases such as feasibility searches and transparent cohort documentation in study publications.

Keywords. formalized cohort definition, FHIR, data requirements

1. Introduction

The application of artificial intelligence (AI) in the medical field has been a rapidly growing area of research, promising significant advancements in diagnostics, treatment optimization, and patient care. However, one of the most critical challenges facing researchers in this domain is the lengthy and resource-intensive process of defining, accessing, and utilizing clinical data.

Despite the pressing need for efficient data handling mechanisms, there is currently no internationally accepted standardized approach for the computer-interpretable representation of study parameters and cohort definitions. Instead, these parameters are

[1] Corresponding Author: Konstantin Knorr, email: knorr@hochschule-trier.de

often conveyed through unstructured text, which can be ambiguous and lacks the consistency required for seamless reuse. This limitation affects various stages of the research process, including study proposals, ethical approvals, data searches, and the eventual publication of results.

Previous studies have attempted to address these challenges through various approaches. For example, [1] analyzed requirements for feasibility queries and developed a Clinical Cohort Definition Language for feasibility queries. Other research highlights the importance of standardized data dictionaries and databases for consistent study comparisons [2], alongside efforts to combine machine learning and rule-based methods to convert eligibility criteria into executable OMOP Common Data Model-compliant queries [3]. Approaches based on standardized frameworks such as FHIR (Fast Healthcare Interoperability Resources) have also been investigated. For instance, systems have been developed to define clinical study inclusion criteria and enable automatic patient searching, particularly focusing on cardiology [4]. Another study presented a method that matches FHIR-based patient data against clinical study inclusion criteria to identify suitable candidates [5].

Despite these advancements, the field still lacks a broadly accepted, structured, and standardized approach to define cohort definitions in a manner that is both user-friendly and intuitively applicable. Therefore, the aim of this study is to propose a novel, technology-agnostic framework for the structured, standardized collection of cohort definitions. This approach seeks to provide a more consistent, reproducible, and efficient method for researchers and healthcare institutions to define data for clinical studies.

2. Methods

We conducted a systematic review of retrospective studies available on ClinicalTrials.gov[2]. All studies were filtered using the search term *Retrospective Study* and the location filter *Germany*, which also includes worldwide studies with at least one participation partner in Germany. Out of the 579 studies meeting these criteria at the time of the search, we randomly selected 164 (≈ 28 %, see [6]) studies and analyzed their eligibility criteria. Different categories reflect distinct aspects of criteria, such as temporal constraints, or clinical conditions when defining cohorts. In the next step, we formalized these categories using a FHIR-compliant format.

3. Results

Six modelling categories were derived from the findings of the study review. We propose a framework that enables the definition of cohort criteria via support for

1. **Patient Demographics** The framework must support the definition of patient data based on demographic characteristics, including age or age ranges, gender, and place of residence (e.g., country and postal code).
2. **Standardized Medical Terminology** Medical concepts within the framework must be defined using standardized coding systems, such as ICD-10/ICD-11, LOINC, and SNOMED CT. Medical concepts must be mappable to multiple codes across different systems simultaneously.

[2] https://clinicaltrials.gov/

3. **Clinical Results Definition** The framework must support the specification of medical result parameters, including measured values (e.g., laboratory results), reference ranges, and clinical interpretations.
4. **Temporal Data Representation** The framework must support the specification of precise time points or intervals of medical events or results.
5. **Temporal Relationships and Dependencies** The framework must support modeling (temporal) dependencies between various medical events including recurring events or results within a specific timeframe (e.g., "Laboratory value X exceeded the threshold at least three times") and time-based dependencies between events or relative to the patient's age (e.g., "Diagnosis X must be made at least six months before Diagnosis Y").
6. **Logical Combination** All definitions must be combinable using logical operators (AND, OR, NOT) enabling cohort definitions.

Our goal is to represent cohort definitions directly within FHIR resources, by using and extending the respective properties in which original information is stored. This approach enhances both human readability for those with FHIR knowledge and the automatic search for data fulfilling the criteria specified in the document. For each category, we state how they can be modelled in a FHIR-compliant format. An example of a complete specification document of a research study including two cohort definitions can be found on GitHub [6].

To implement the **Patient Demographics** category, we first introduce a new FHIR object that represents an operator/value-pair. We extend properties 'birthDate', 'gender' and 'address.postalCode' of the 'Patient' resource to optionally accept such operator/value-pairs. Accordingly, we realize **Standardized Medical Terminology** by extending the 'code' property of resources like 'Observation' or 'Condition' to also accept operator/value-pairs. Since 'coding' is an array of objects, we allow multiple entries. Each entry then represents a different code system and all entries are connected via logical OR.

```
{
  "coding": [{
    "system": < ICD-10 | LOINC | SNOMED CT | OPS | ATC |
      ... >,
     "code": {
         "operator": < "$eq" | "$ne" | "$in" | "$nin" >
               "value": < string | list >
       }
   }]
}
```

The **Clinical Results Definition** does not require any extension, it can be captured by existing FHIR properties. To define measured values, we use the 'valueQuantity' property of the 'Observation' resource type. This property allows to specify a measured value, its unit and additionally a quantity comparator. To define ranges of measured values, we use the 'valueRange' property; clinical interpretations are stored in the 'interpretation' property of 'Observation'.

Temporal Data Representation can be captured using operator/value-pairs if we want to express that a medical event took place before, after, or on a particular date or time span. We hence extend the '<onset | effective | performed | ... > dateTime' properties of aforementioned resource types accordingly. To determine the *duration* of an event, '<onset | effective | performed | ... > Period' properties are used. Therefore, we extend the operator/value-pair from above by the following comparator object:

```
{
   "LHSvalue":  < comp. value | computation >,
    // e.g., perfomedPeriod.end - performedPeriod.start
    "operator": < "$lt"|"$lte"|"$gte"|"$gt"|"$eq"|"$ne" >,
    "RHSvalue": < comparison value >
    // e.g., 5 days
}
```

We use the 'Encounter' resource to model **Temporal Relationships and Dependencies**. We assume that medical events and observations are associated with a specific encounter. To explicitly state that certain events must have occurred during the same encounter, we use the encounter property to reference a shared encounter instance.

If medical events shall be linked, but do not necessarily need to be related to the same encounter, we extend the 'basedOn' array of the respective resource. Each element in this array is enhanced to accept references to 'Encounter' and 'Patient', along with an optional comparison object to express more complex dependencies. These comparisons refer to properties of the linked 'Encounter' ('Patient') via their respective referenced IDs. Referenced encounters may include additional requirements, such as their encounter type, or observations and conditions that must have occurred. All entries within the basedOn array are combined using logical AND, meaning all specified conditions must be satisfied.

Throughout this modelling and even if not directly specified, it is implicitly understood that all referenced data pertains to the same patient. This approach is not limited to the 'Encounter' resource; any FHIR resource that includes the 'basedOn' property can be extended in the same manner to express dependencies or comparisons between, for example, medications.

```
"basedOn": [{
   "reference": "Encounter/enc2",
   "LHSvalue":  < computation >,
   // e.g., enc2.cond2.onsetDateTime -
enc1.cond1.onsetDateTime
   "operator": < "$lt"|"$lte"|"$gte"|"$gt"|"$eq"|"$ne" >,
   "RHSvalue": < comparison value >
}]
```

Logical Combinations are made as follows: An entire cohort definition is made by adding all references of all entries to a 'Group' resource that is referenced in the 'enrollment' array of 'ResearchStudy'. Within each group, all requirements are 'AND'-connected. If multiple cohorts need to be defined, additional groups can be added that are considered independently of each other.

4. Discussion and Conclusion

In this paper, we proposed a framework for the structured and standardized specification of cohort definitions for retrospective medical research. From a systematic review of studies, we derived six modelling categories representing key criteria for defining study populations, which we directly embedded into FHIR resources, ensuring a one-to-one correspondence between criteria definition and original data. The resulting specification is both human- and machine-readable, and can be used for an automated search of existing data as described in [7] or for a transparent data description in the publication of a study. A user-friendly interface facilitates the formalization process without needing to have knowledge of the FHIR standard, as proposed in [8].

To enable data search using the proposed specification document, data must be collected and stored in a universally standardized way, using established standards, terminologies, and ontologies, for example FHIR or ICD. However, this level of standardization is not guaranteed in practice, as hospitals often use institution-specific coding systems and data formats. While many data requirements can be mapped to our six proposed modelling categories, certain limitations remain. The current framework lacks the ability to specify follow-up studies of specific previous studies, include planned or optional procedures, or define requirements for family history or anamnesis data. Although the framework cannot yet capture all types of study requirements, we consider it a meaningful step toward standardizing retrospective study definitions.

Acknowledgement

This work is part of the DaTreFo project, funded by the German Federal Ministry of Research, Technology, and Space (16KIS1644).

References

[1] Rosenau L, Gruendner J, Kiel A, Köhler T, Schaffer B, Majeed RW, et al. Bridging Data Models in Health Care With a Novel Intermediate Query Format for Feasibility Queries: Mixed Methods Study. JMIR medical informatics. 2024;12(1):e58541. doi:10.2196/58541.

[2] Shenvi EC, Meeker D, Boxwala AA. Understanding data requirements of retrospective studies. International Journal of Medical Informatics. 2015 Jan;84(1):76-84. doi:10.1016/j.ijmedinf.2014.10.004.

[3] Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. Journal of the American Medical Informatics Association. 2019;26(4):294-305. doi:10.1093/jamia/ocy178.

[4] Scherer C, Endres S, Orban M, Kääb S, Massberg S, Winter A, et al. Implementation of a clinical trial recruitment support system based on fast healthcare interoperability resources (FHIR) in a cardiology department. European Heart Journal-Digital Health. 2022;3(4):ztac076-2795. doi:10.1093/eurheartj/ehac544.2795.

[5] Alper BS, Dehnbostel J, Shahin K, Ojha N, Khanna G, Tignanelli CJ. Striking a match between FHIR-based patient data and FHIR-based eligibility criteria. Learning Health Systems. 2023 Oct;7(4):e10368. doi:10.1002/lrh2.10368.

[6] Supplement Material to the Paper; https://github.com/CaroPoschen/SupplementMaterial/.

[7] Herres B, Poschen C, Knorr K. Privacy-Preserving Search on Medical Data. In: Digital Health and Informatics Innovations for Sustainable Health Care Systems. IOS Press; 2024. p. 252-6. doi:10.3233/SHTI240392.

[8] Poschen C, Berens B, Knorr K. Towards Formalized Study Parameters for Medical Research; 2025. To be published at 17th International Conference on e-Health, 2025.