Bridging Data Models in Healthcare: Introducing a novel intermediate query format for feasibility queries.

Lorenz Rosenau, Julian Gruendner, Alexander Kiel, Thomas Köhler, Bastian Schaffer, Raphael Majeed

Submitted to: JMIR Medical Informatics
on: March 18, 2024

Table of Contents

Bridging Data Models in Healthcare: Introducing a novel intermediate query format for feasibility queries.

Lorenz Rosenau[1*] MSc; Julian Gruendner[2*] PhD; Alexander Kiel[3] BSc; Thomas Köhler[4, 5, 6]; Bastian Schaffer[2]; Raphael Majeed[7] PhD

[1]IT Center for Clinical Research University of Lübeck Lübeck DE
[2]Chair for Medical Informatics Friedrich-Alexander-Universität Erlangen-Nürnberg Erlangen DE
[3]Leipzig Research Centre for Civilization Diseases University of Leipzig Leipzig DE
[4]Faculty of Medicine Mannheim University Heidelberg Mannheim DE
[5]German Cancer Consortium Heidelberg DE
[6]German Cancer Research Center Heidelberg DE
[7]Institute for Medical Informatics University Clinic Rheinisch-Westfälische Technische Hochschule Aachen Aachen DE
[*]these authors contributed equally

Corresponding Author:
Lorenz Rosenau MSc
IT Center for Clinical Research
University of Lübeck
Gebäude 64, 2.OG, Raum 05
Ratzeburger Allee 160
Lübeck
DE

Abstract

Background: To advance research with clinical data, it is essential to provide researchers with fast and easy access to the available data, which is especially challenging for data from different source systems within and across institutions. Over the years, many research repositories and data standards have been created. One of these is the FHIR standard, used by the German Medical Informatics Initiative (MII) to harmonize and standardize data across university hospitals in Germany. One of the first steps to make this data available is to allow researchers to create feasibility queries to determine the data availability for a specific study. Given the heterogeneity of different query languages to access different data across and even within standards such as FHIR (e.g., CQL and FHIR Search), creating an intermediate query syntax for feasibility queries reduces the complexity of query translation and improves interoperability across different research repositories and query languages.

Objective: This study describes the creation and implementation of an intermediate query syntax for feasibility queries and how it integrates into the federated German health research portal (Forschungsdatenportal Gesundheit) and the German medical informatics initiative.

Methods: We analyzed the requirements for feasibility queries, and the currently available feasibility tools that research repositories offer. Based on this analysis, we developed an intermediate query syntax that can be easily translated into different research repository-specific query languages.

Results: The resulting Clinical Cohort Definition Language (CCDL) for feasibility queries combines inclusion criteria in a conjunctive normal form and exclusion criteria in a disjunctive normal form, allowing for additional filters like time or numerical restrictions. The inclusion and exclusion results are combined via an AND NOT expression to specify complex feasibility queries. We defined a JSON schema for the CCDL, generated an ontology, and demonstrated the use and translatability of the CCDL across multiple studies and real-world use cases.

Conclusions: We developed and evaluated a structured query syntax for feasibility queries and demonstrated its use in a real-world example as part of a research platform across 39 German university hospitals.

(JMIR Preprints 18/03/2024:58541)
DOI: https://doi.org/10.2196/preprints.58541

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Lorenz Rosenau[a], Julian Gruendner[b], Alexander Kiel[c], Thomas Köhler[d,e,f],
Bastian Schaffer[b], Raphael Majeed[g]

[a]IT Center for Clinical Research, University of Lübeck, Lübeck, Germany
[b]Chair for Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[c]Leipzig Research Centre for Civilization Diseases, University of Leipzig, Leipzig, Germany
[d]German Cancer Research Center, Heidelberg, Germany
[e]Faculty of Medicine Mannheim, University Heidelberg, Mannheim, Germany
[f]German Cancer Consortium, Heidelberg, Germany
[g]Institute for Medical Informatics, University Clinic Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany

Bridging Data Models in Healthcare: Introducing a novel intermediate query format for feasibility queries.

Keywords: Feasibility, FHIR, CQL, Eligibility Criteria, Clinical Research, Intermediate Query Format, Healthcare interoperability

Abstract

Background: To advance research with clinical data, it is essential to make access to the available data as fast and easy as possible for researchers, which is especially challenging for data from different source systems within and across institutions. Over the years, many research repositories and data standards have been created. One of these is the FHIR standard, used by the German Medical Informatics Initiative (MII) to harmonize and standardize data across university hospitals in Germany. One of the first steps to make this data available is to allow researchers to create feasibility queries to determine the data availability for a specific research question. Given the heterogeneity of different query languages to access different data across and even within standards such as FHIR (e.g., CQL and FHIR Search), creating an intermediate query syntax for feasibility queries reduces the complexity of query translation and improves interoperability across different research repositories and query languages.

Objective: This study describes the creation and implementation of an intermediate query syntax for feasibility queries and how it integrates into the federated German health research portal (Forschungsdatenportal Gesundheit) and the MII.

Methods: We analyzed the requirements for feasibility queries, and the feasibility tools that are currently available in research repositories. Based on this analysis, we developed an intermediate query syntax that can be easily translated into different research repository-specific query languages.

Results: The resulting Clinical Cohort Definition Language (CCDL) for feasibility queries combines inclusion criteria in a conjunctive normal form and exclusion criteria in a disjunctive normal form, allowing for additional filters like time or numerical restrictions. The inclusion and exclusion results are combined via an AND NOT expression to specify feasibility queries. We defined a JSON schema for the CCDL, generated an ontology, and demonstrated the use and translatability of the CCDL across multiple studies and real-world use cases.

Conclusions: We developed and evaluated a structured query syntax for feasibility queries and demonstrated its use in a real-world example as part of a research platform across 39 German

university hospitals.

Introduction

In the rapidly evolving field of medical research, patient data has emerged as a critical resource. The vast amounts of data generated through clinical encounters, laboratory tests, imaging studies, and other patient interactions hold the potential to significantly advance our understanding of disease processes and treatment outcomes. Clinical Data Repositories (CDRs) are a valuable tool for storing, organizing, and retrieving this wealth of patient data. These repositories facilitate data storage in a structured and standardized manner, enabling researchers to query this data efficiently for various research purposes.

One key aspect of effectively utilizing CDRs is the ability to perform feasibility queries. These queries allow researchers to assess the availability and adequacy of data for specific research questions before embarking on full-scale studies. Doing so can save considerable time and resources by identifying potential issues, such as insufficient sample size or a lack of necessary data elements.

Distributed data collections

The landscape of data repositories is not homogenous. There are two primary approaches to data repository management: the classical single repository approach and the federated approach. Traditionally, these repositories have been centralized, pooling data from various sources into a single repository [1]. However, this classical approach has been challenged by the emergence of federated data repositories [1,2].

The classic single repository approach involves a centralized system where all data is stored and managed in one place. This solution offers the advantage of uniformity and ease of data management. It enables efficient data quality benchmarking at scale and the generation of derivatives, harmonized variables, and units of measure for comparable and consistent analytics [1]. However, it is often impractical or impossible to implement, especially when dealing with multiple institutions, each having its own schema for its clinical data repository.

On the other hand, the federated approach involves a network of repositories, each maintained by different institutions. These repositories operate independently but are interconnected for data sharing and collaboration. The data generally remains at the generating site, which offers the advantages of local curation by personnel deeply familiar with the data [1] and maintains data anonymity and security [2]. The data can then be analyzed using a federated approach or, if the correct patient consent is given, be transferred to a central data management unit for a specific analysis.

This approach respects individual institutions' autonomy and data governance policies, making it a more feasible option for multi-institutional collaborations [3–9] and can enhance the scope and depth of clinical research by enabling access to a broader range of data.

Despite the potential benefits of federated data repositories, performing feasibility queries across multiple CDRs presents significant challenges [10]. Each repository contains data originating from different source systems, leading to heterogeneity in data formats, terminologies, and quality. This heterogeneity can significantly complicate the process of data integration and harmonization, making it challenging to perform comprehensive and accurate feasibility queries [10].

Moreover, the federated nature of the system introduces additional complexities. Data privacy regulations and institutional policies may restrict the sharing and use of certain data, further complicating the query process. Additionally, the technical infrastructure required to support secure and efficient data exchange across multiple repositories can be challenging to implement and maintain.

Data exchange standards for interoperability

In a federated network, the commitment to an interoperability standard becomes pivotal to tackling

these challenges. Prominent examples include but are not limited to FHIR [11], OMOP CDM [12], i2b2 [13], and OpenEHR [14] share the commonality of being centered around the patients' medical history.
Agreeing on an interoperability standard only partially solves the challenge. While a healthcare data exchange standard facilitates the conversion of existing data into a common format at each hospital, a distributed feasibility query platform for the data is still missing.

Tools for feasibility queries

Besides the data integration standardization, interactive user interfaces enable researchers to design and submit feasibility queries. For this purpose, a multitude of tools for feasibility queries exist (e.g., i2b2 [13,15], TriNetX [16], tranSMART [17], SampleLocator [18–20], Observational Health Data Science and Informatics (OHDSI) ATLAS [21], DZHK Feasibility Explorer [22]), each with its own data formats, standards, and query languages, including Structured Query Language (SQL), Clinical Quality Language (CQL), FHIR-Search, and Archetype Query Language (AQL). Consequently, querying across these different tools is difficult as there is no common query representation, and researchers must navigate these diverse tools and formats, particularly when dealing with cross-institutional data or distributed data storage within an institution.
Within the broader context of establishing a feasibility platform as part of the central German Portal for Health Data (FDPG), this research introduces a novel query syntax, serving as an intermediary between user interfaces and data repositories. This syntax is designed to be sufficiently flexible to ensure interoperability while maintaining simplicity. It focusses on the primary needs of a feasibility query, while allowing the syntax to be translated into repository-specific languages like FHIR-Search or CQL.

Our approach is grounded in the broader context of clinical research, where the reuse of eligibility criteria is common, whether in their original form or with modifications. These criteria are instrumental not just for feasibility studies but also for pre-screening, data selection, extraction, and validation. Consequently, a need has emerged to decouple the representation of eligibility criteria from their implementation in specific systems. A mechanism to express complex criteria and combinations thereof in a way that is both intuitive and adaptable to varying implementation needs is required.

In this study, we describe the development and application of the query syntax within the network of the Medical Informatics Initiative (MII), encompassing 39 German university hospitals, specifically, the FDPG feasibility platform and show how it achieves interoperability across different research platforms.

Methods

Requirement analysis

In our pursuit to create an intermediate query syntax to express eligibility criteria, we performed a requirement analysis. Within it, we combine insight from feasibility queries and cohort selection, with the latter often manifesting as a query output in the form of cohort size rather than a list of discrete patient identifiers.
Our research reviewed existing tooling, namely i2b2, TriNetX, and OHDSI Atlas. We aimed to identify common functionalities and essential features across these tools. To obtain insight into the criteria's structure and complexity, we analyzed existing eligibility criteria from ClinicalTrials.gov [23] and incorporated the findings from Ross et al. [24] and Gulden et.al. [25]. Moreover, we integrated insights from the usability study by Schüttler et al. evaluating feasibility tools [26], conducted expert interviews and recursively synchronized the requirements within our project

(https://pubmed.ncbi.nlm.nih.gov/35486893/). This multifaceted analysis allowed us to infer a set of requirements crucial for developing our query syntax. These requirements were categorized into query expressiveness, interoperability, and accessibility.

Expressiveness requirements

The query syntax should:

- Allow for the definition of inclusion and exclusion criteria
- Be expressed in Boolean logic
- Allow the expression of exclusion criteria
- Support at least patient as query subject (feasibility queries can be performed on different query subjects: find all patients with specific criteria, find all encounters with specific criteria, find all specimens with specific criteria)
- Use unique identifiers for criteria and concepts
- Support the following filter on the criterion level:
  - existence of a criterion
  - numeric restriction
  - concept filter
  - time restrictions
  - attribute filters

Interoperability and accessibility requirements

The query syntax should:

- Provide an abstract (decoupling) layer between the user interface and the query execution.
- Have a low level of complexity and be easily translatable to different query languages.
- Be suitable for integration with the HL7 FHIR standard used by the MII.
- Use a widely used data exchange format like JSON to ease parsing and generation
- Human readability/writability
- Ideally directly support the use of standard medical terminology (LOINC, SNOMED-CT, ICD10, …) to lower mapping efforts

Related work and existing solutions

Analyzing the existing solutions, we found that none of the solutions met all the requirements. Most failed to have a formally defined low-complexity feasibility query syntax, and i2b2 was missing the direct relationship with the terminology on the syntax level. FHIR Search and the FHIR standard did not provide the ability to express a feasibility query in the required scope [25] at the time of our research. Other query languages that could have been candidates, like CQL or SQL, are complex or data model specific, making the translation between different data models and their representation, as well as the generation of the syntax by a user interface, challenging.

Evaluation

To evaluate the specification of the query syntax, we compared the final specification with our requirements and additionally demonstrated its applicability beyond the scope of FHIR by applying it to AQL.

We incorporated the solution into a large-scale real-world distributed feasibility query infrastructure, including a user interface, where it was integrated as the central intermediate query syntax. We further evaluated the applicability of the syntax to a wide range of clinical criteria and investigated its translatability, as well as how well it lends itself to creating a user interface for feasibility queries. Beyond the use in our projects based on German datasets and specifications, we also successfully applied the CCDL to the international Synthea [27] dataset.

Results

Based on the requirements of a team of experts, we created the “Clinical Cohort Definition Language” (CCDL), an intermediate query syntax for feasibility queries. The exchange format for the syntax was chosen to be JSON, which is currently widely used across the software community and is familiar to software developers from user interfaces, REST APIs, and query execution backends alike.

Criterion types and filters

The atomic component of CCDL is the criterion, serving as the foundational building block for inclusion or exclusion criteria. Each criterion is uniquely identified using a tuple of code system and code (which we named termCode) analogous to FHIR and OMOP-CDM (For conceptual equivalence between concepts across medical terminologies, multiple termCodes can be provided, e.g., the criterion for sleep apnea may be represented by the termCodes G47.3 from ICD-10 and 73430006 from SNOMED CT). Each termCode may have an additional “display” attribute, which serves purely as a visual representation to make the interpretation of a CCDL easier for humans. Within our CCDL, the criteria can occur as one of four different base types of criteria:

- **Exist criteria** with no additional filters (e.g., conditions or a laboratory concept with no filter, like the existence of a Hemoglobin value regardless of the value)
- **Comparatively restricted numerical criteria** (e.g., Hemoglobin laboratory value < 12 g/dL)
- **Range-restricted numerical criteria** (e.g., Hemoglobin laboratory value between 10 and 12 g/dL)
- **Value set restricted criteria** (e.g., gender = female, male)

Additionally, each criterion can be further restricted to a date range (e.g., a Condition that occurred between 01.01.2024 and 05.02.2024), and it supports additional "attribute" filters, which can be added to each of the base types of criteria. The attribute filters support similar filters that identify the criterion types, i.e., comparative numerical, comparative range, and value set restriction (e.g., the body site = skin for a tissue specimen – see Appendix 1).

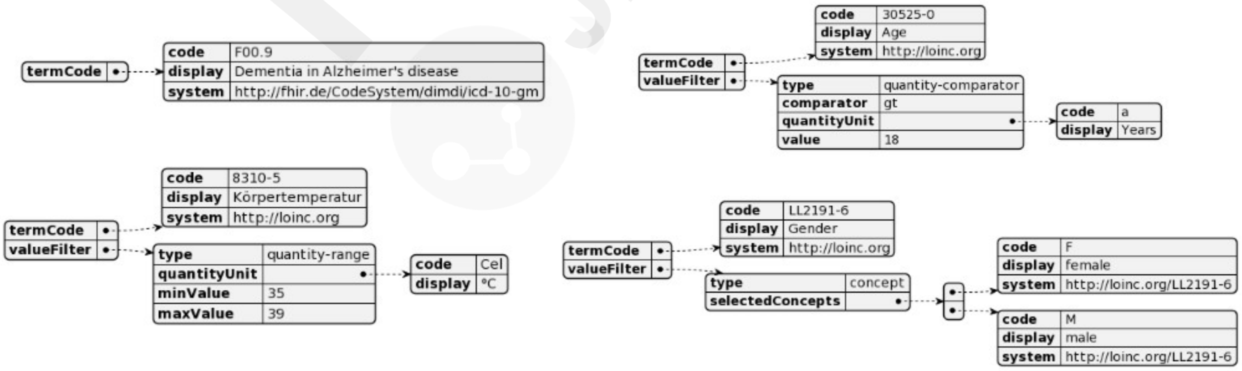


Figure 1: Different types of criteria definitions. A: Simple conceptual criterion. B: Numeric criterion with quantitative comparison. C: Numeric criterion with range restriction. D: Valueset criterion.

The explicit logic layer

The logic layer of the query aligns with existing solutions (i2b2/tranSMART/TriNetX) in representing the structured query as a combination of conjunctive normal form (CNF) and disjunctive normal form (DNF). Every criterion is embedded into the logic layer in a CNF for inclusion criteria and DNF for exclusion criteria (Figure 2). Inclusion and Exclusion criteria are then logically combined via an AND NOT operator by subtracting the result of the exclusion criteria from the result of the inclusion criteria. Every feasibility query also receives a syntax version number and an additional description. The syntax version allows to distinguish the current version from future versions and changes, and the description allows the query to transport additional human-readable information about the query.

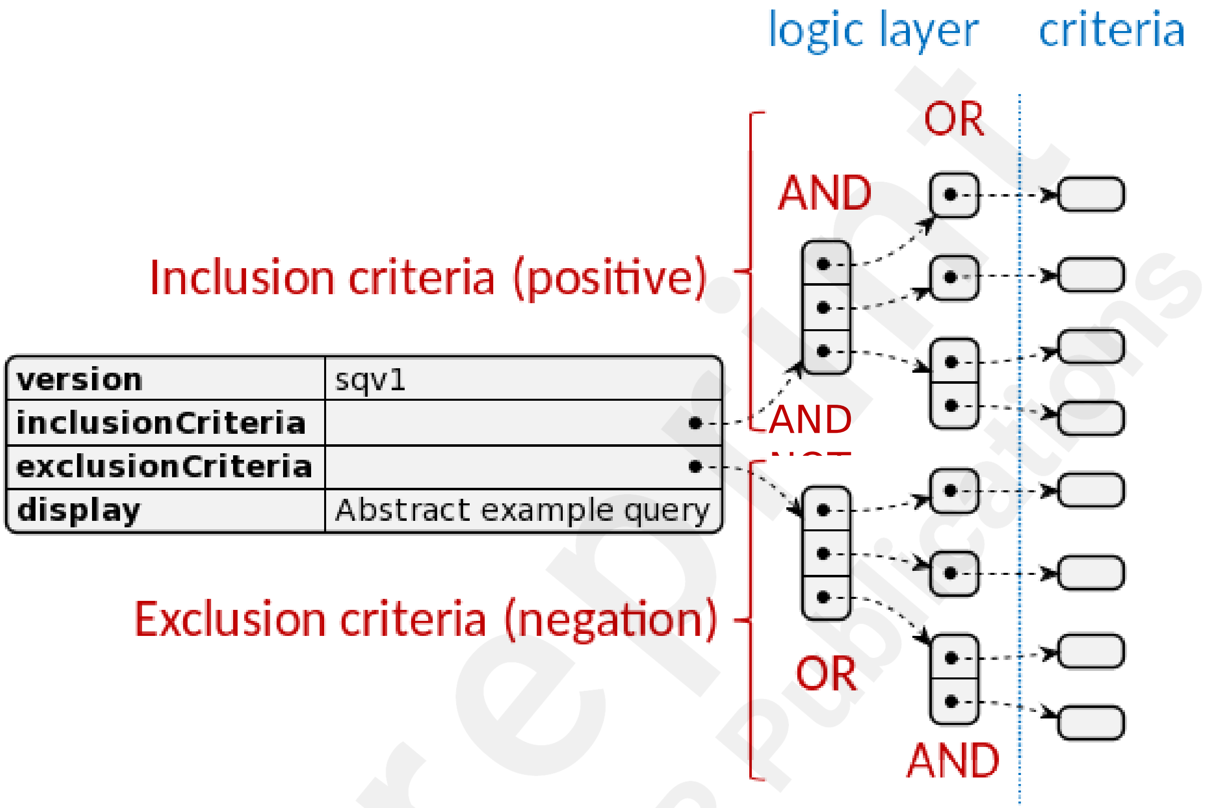


Figure 2. Structured query syntax top-level elements and logic layer. Certain criterion types will imply additional intrinsic logical relations. See Valueset-Criteria and attribute filters and time restrictions.

The implicit logic (criteria expansion)

Apart from the explicit logic layer across criteria, different types of criteria and their filters further impact the execution logic as follows.

ValueSet criteria (see Figure 1 D) allow the selection of multiple values (concepts). In this case, the value selections are treated as OR choices. For example, gender = (male, female) expands to: (gender=male) OR (gender=female).

Attribute filters for each criterion are additional filters that can be set for each criterion. All individual filters on a criterion are combined using AND. For example, a specimen of type "Tissue specimen" and body site "skin" only applies to specimens with the type of Tissue and the body site skin.

The same applies to time restrictions. In this example, the time restriction "between 2020-01-01 and 2021-01-01" will predictably be added using an AND conjunction of the type, body site, and time restriction.

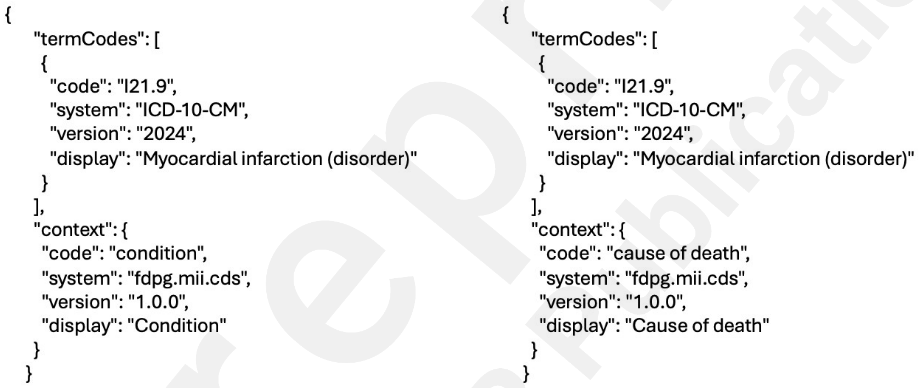
Furthermore, there is an implicit OR expansion of criteria when the criterion-identifying code is a parent code of multiple child codes within a terminology hierarchy. For example, suppose a researcher adds the diagnosis of Diabetes mellitus, Type 2, as a criterion (ICD10 code=E11). In that case, it can be expanded to search all subtypes of Diabetes mellitus, Type 2 (E11, E11.3, E11.31, E11.30, E11.1, E11.11, E11.0, E11.01, E11.7, E11.75, E11.74, E11.73, E11.72, E11.4, E11.41, E11.40, E11.8, E11.81, E11.80, E11.2, E11.21, E11.20, E11.5, E11.51, E11.50, E11.6, E11.61, E11.60, E11.9, E11.91, E11.90) combining them using a logical OR operation).

Context-dependent criteria

In some cases, a criterion cannot be uniquely defined by its term code within a terminology, making it impossible to map a criterion for execution. One example of this is the use of ICD-10 condition codes for causes of death, specimen-specific conditions, or the general condition of a patient.

In modern terminologies like SNOMED-CT, this can be resolved using post-coordination, where a combined code, which carries the context, is created. For example, 419620001|Death|:42752001|Due to|=22298006|Myocardial infarction| which, while in line with SNOMED Compositional Grammar [28], a template to express this is not currently part of the SNOMED CT implementation.

The syntax we developed here allows for post-coordinated codes; however, we allow for an additional "context" attribute for some use cases where post-coordination is unsuitable. The context attribute is modeled after our termCode attribute and provides an extra term code to identify the context. Figure 3 provides an example for Myocardial infarction as condition aand cause of death.



```
{
  "termCodes": [
   {
    "code": "I21.9",
    "system": "ICD-10-CM",
    "version": "2024",
    "display": "Myocardial infarction (disorder)"
   }
  ],
  "context": {
   "code": "condition",
   "system": "fdpg.mii.cds",
   "version": "1.0.0",
   "display": "Condition"
  }
 }
```

```
{
  "termCodes": [
   {
    "code": "I21.9",
    "system": "ICD-10-CM",
    "version": "2024",
    "display": "Myocardial infarction (disorder)"
   }
  ],
  "context": {
   "code": "cause of death",
   "system": "fdpg.mii.cds",
   "version": "1.0.0",
   "display": "Cause of death"
  }
 }
```

Figure 3: Myocardial infarction in two contexts (condition and cause of death)

Data availability

As a technical solution to define the structure of the CCDL, we decided on the JSON Schema definition and made it publicly available [29]. The schema serves implementation guidance and validation purposes; the git repository also contains documentation examples, test data, and the capabilities to create matching test queries.

Requirement verification

An analysis was performed based on the structure defined in the JSON schema to evaluate the developed intermediate query syntax. The following table presents the detailed results of this analysis:

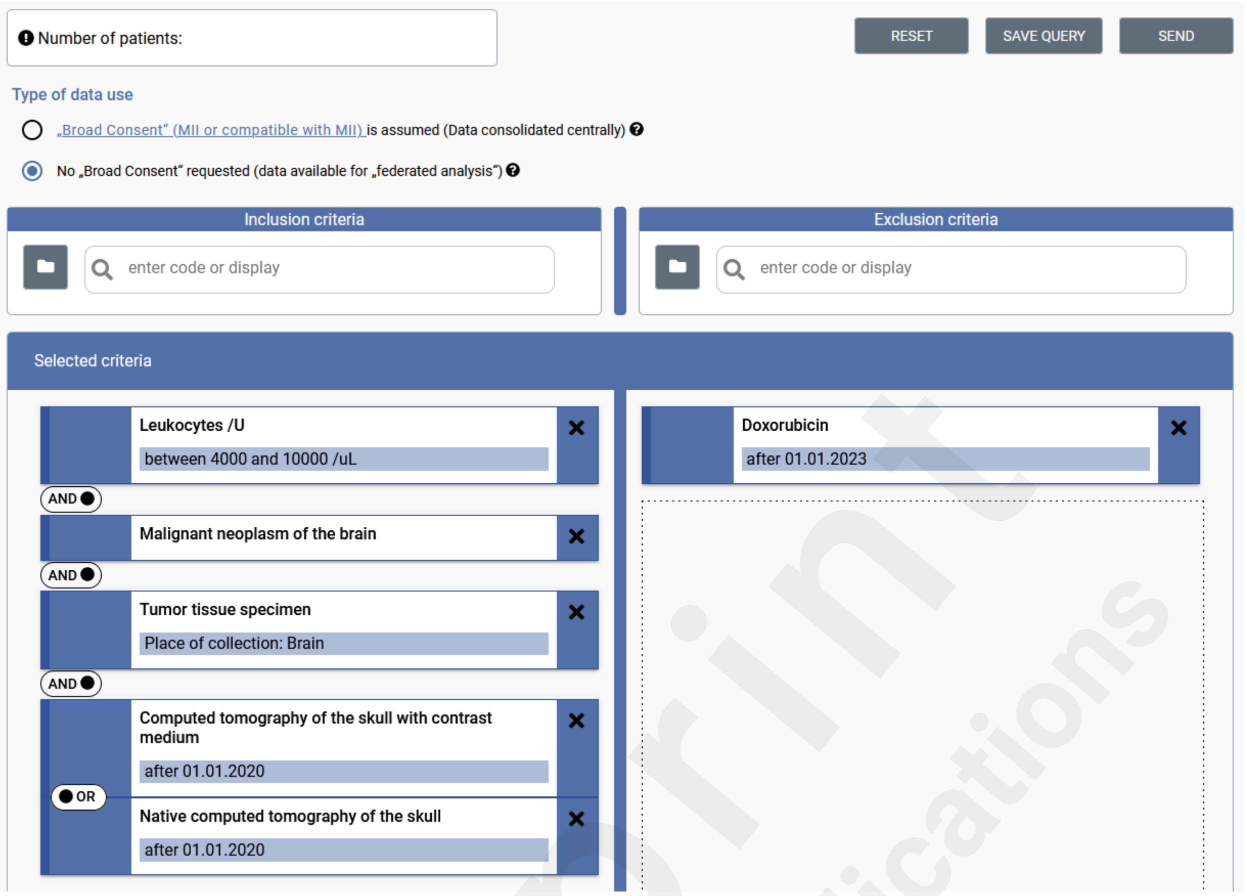
Component	Key Properties	Purpose & Function	Requirements Met
inclusionCriteria	CNF without negation	Conjunction of criteria with logical operators.	Expressive query formulation, boolean logic
exclusionCriteria	DNF without negation	Allows negation of criteria for comprehensive exclusion.	Negation of criteria on a group level, Boolean logic
termCode	code, system, version, display	Identifies concepts using standard coding systems.	Standard medical terminology, uniqueness
criterion	context, termCodes, valueFilter, attibuteFilter, timeRestriction	Sets criteria with defined context, using term codes and filters.	Expressiveness of simple and complex eligibility criteria
timeRestriction	afterDate, beforeDate	Specifies time frame for criteria fulfillment.	Time restrictions
unit	code, display	Standardized unit definition, adhering to UCUM units.	Use of standardized units
valueFilter	type (concept, quantity-comparator, etc.)	Varied filtering types for flexible data querying.	Numeric restriction, concept restrictions
attributeFilters	type (concept, quantity-comparator, reference)	Mechanism for detailed filtering at the attribute level.	Detailed filtering, clinical relations

Table 1: CCDL Components and their purpose regarding the expressiveness requirements

This syntax efficiently meets a wide range of expressiveness and interoperability requirements, demonstrating capabilities in defining complex medical queries with standard terminologies and logical operators.

Evaluation and use of the CCDL in real-world scenarios

We believe the potential of the Clinical Cohort Definition Language (CCDL) extends beyond its application in the federated feasibility portal of the German Research Portal for Health. Nevertheless, the CCDL remains a crucial technical solution within the FDPG's feasibility portal.



Number of patients:

Type of data use

○ „Broad Consent" (MII or compatible with MII) is assumed (Data consolidated centrally)

● No „Broad Consent" requested (data available for „federated analysis")

Leukocytes /U
between 4000 and 10000 /uL

AND

Malignant neoplasm of the brain

AND

Tumor tissue specimen
Place of collection: Brain

AND

Computed tomography of the skull with contrast medium
after 01.01.2020

OR

Native computed tomography of the skull
after 01.01.2020

Doxorubicin
after 01.01.2023

Figure 4: Example of a feasibility query in the FDPG Feasibility Portal to find patients with a leukocyte count within a normal range, with a malignant neoplasm of the brain, available tumor tissue specimen, and a CT scan after 01.01.2020 who did not take Doxorubicin.

We created a user interface for the feasibility queries in the FDPG, which generates the CCDL, demonstrating how it lends itself well to building feasibility query user interfaces [30,31] The CCDL especially supports this as its design follows the typical way of feasibility query creation as seen in platforms such as the FDPG, i2b2, OMOP, and TrinetX. We evaluated the usability of the user interface across multiple projects [32,33] and embedded it in a German-wide distributed research infrastructure [10,34]. These evaluations highlighted the applicability of the CCDL to a feasibility query and that the usability issues found were not due to a lack of expressiveness of the CCDL. We further used the Synthea dataset to test the CCDL against [35], demonstrated the ability of the CCDL to represent a wide range of criteria [36], and showed that it could be fully translated to FHIR Search [37], CQL [38], and AQL [39]. At the time of writing, almost 9000 CCDLs have been created and executed across Germany.

Discussion

Principal Findings

We presented an intermediate feasibility query syntax that separates concerns between the user interface and the execution of a feasibility query on different research repositories and their specific query languages. The syntax defined here fulfills all the interoperability and accessibility requirements while supporting a broad range of expressiveness requirements we identified by analyzing existing query tools. The solution is fully compatible with established medical terminology standards, notation of parameters, and restriction semantics.

The solution we describe here is compatible with the query logic established by i2b2 and, therefore, tranSMART and TriNetX. This means that tools like i2b2 or similar could be easily extended to produce our syntax.

The CCDL was further used as part of a larger infrastructure for feasibility queries in Germany and is currently used as the interface for feasibility queries within the German research portal for health, supporting feasibility queries across 39 university hospitals in Germany. We successfully created translation components for FHIR Search and CQL in the current implementation. Current research also indicates the adaptability for FHIR Pathling's aggregation API [40], and SQL. The criteria content and the required reintroduction of data model-dependent information are obtained from an automatically generated search ontology [41].

Related Work

While the expression of eligibility criteria within a specific data model context is well established and adequately discussed in this work, research on a data-agnostic intermediate format for computable eligibility has been sparse in recent years.

Alper et al. [42] closely align with our approach of representing eligibility criteria in a structured format, namely the FHIR EvidenceVariable, which currently does not directly support the representation of eligibility criteria but may be refined to do so. Presumably, a FHIR representation would provide a structure beyond the realm of the MII, which could add significant value and improve syntax interoperability. However, in the early stages, the challenges of adopting new solutions could have impeded the development presented here. Our ongoing communication with the HL7 (Health Level Seven International) working group, which focuses on Research Studies, gives us confidence that once a suitable FHIR Resource is established or adapted to meet the needs outlined in our publication, the established technical components could be efficiently modified to align with these changes. Parallels can be drawn to implementing structured eligibility criteria, as presented by Fang and Yuan et al. [43,44]. Their publications present a half-automated approach to generate feasibility queries based on free text study protocols from ClinicalTrials.gov [23] Their system is built around the OHDSI data model and uses the concept IDs. After converting the free text criteria, they allow users to edit and download an intermediate representation in JSON format. Unfortunately, no clear implementation guidelines on the format are given by Fang and Yuan et al. However, recurring themes include differentiating inclusion and exclusion criteria and defining temporal constraints. To our knowledge, contrary to our approach, they do not allow for further restrictions beyond the value constraint on specific criteria.

Limitations

The separation of concerns, which the CCDL provides, also leads to the need for a mapping to identify the correct way of translating the CCDL information model to the local information model and terminology. The mapping allows the link between the specific data model and the criterion as identified in the CCDL to be created. One example of this is that for FHIR Search, the mapping for a condition criterion identified by a specific ICD10 code C50.0 would provide the information that the condition is found in the endpoint "/Condition" and the search parameter for the term code is "code" – Leading to the translated FHIR Search URL:
[fhir-base-url]/Condition?code=http://fhir.de/CodeSystem/bfarm/icd-10-gm|C50.0".

Further, additional information about the terminology is necessary to allow the selection of criteria within a terminology hierarchy, where the criterion resolves to multiple child criteria. Finally, this then requires the query executor and the CCDL generating user interface to agree on criteria or terminology entries.

One common requirement currently not supported by the CCDL is temporal interdependencies between different criteria. Therefore, queries like a specific laboratory value within a certain period of diagnosis cannot be currently expressed using the CCDL.

We deliberately decided to delay the implementation of this extension as time dependencies significantly increase the complexity and performance requirements of any query execution.

The data model agnostic nature of the CCDL is inherently valuable. Its full potential — the capability to be utilized across different healthcare data models — requires more than technical translation. For cross-model query capability, the existence of the concepts in all target data models must be ensured.

Future Work

The CCDL described here provides a good base to make feasibility queries possible across various research repositories and close the gap between the different research repositories and their access. We have demonstrated the applicability of the CCDL to FHIR Search, CQL, and AQL; however, more repositories and other query languages, such as SQL on FHIR, OHDSI OMOP, or i2b2 might be added in the future. Further, one could imagine how separating the query syntax and execution would theoretically allow one to query different internationally distributed repositories such as FHIR, OMOP-CDM, and i2b2 simultaneously. Additionally, the CCDL is currently limited in how much it can express, and new capabilities will be added in the future. In this pursuit of making the CCDL more expressive, any extension must be weighed against the added complexity and overhead it introduces.

Conclusion

We presented a query syntax for medical feasibility queries, which creates an abstract layer between the user interface and the execution query language. We showed how it is flexible enough to be translated into different query languages and can be used to express various complex feasibility queries. The applicability of the query syntax was further demonstrated by embedding it into a large research project where it is used to query multiple millions of patients across 39 German university hospitals. The CCDL for feasibility queries will be extended in the future to allow more features, and we are currently working on a modified version for data selection and extraction.

Acknowledgment

The project was funded by the German Federal Ministry of Education and Research (BMBF) under the FDPG-PLUS Project, with grant numbers 01ZZ2309A, 01ZZ2309C, 01ZZ2309D, 01ZZ2309E, 01ZZ2309F.

Conflicts of Interest

None declared.

Literature

1. Pfaff ER, Girvin AT, Gabriel DL, Kostka K, Morris M, Palchuk MB, Lehmann HP, Amor B, Bissell M, Bradwell KR, Gold S, Hong SS, Loomba J, Manna A, McMurry JA, Niehaus E, Qureshi N, Walden A, Zhang XT, Zhu RL, Moffitt RA, Haendel MA, Chute CG, The N3C Consortium, Adams WG, Al-Shukri S, Anzalone A, Baghal A, Bennett TD, Bernstam EV, Bernstam EV, Bissell MM, Bush B, Campion TR, Castro V, Chang J, Chaudhari DD, Chen W, Chu S, Cimino JJ, Crandall KA, Crooks M, Davies SJD, DiPalazzo J, Dorr D, Eckrich D, Eltinge SE, Fort DG, Golovko G, Gupta S, Haendel MA, Hajagos JG, Hanauer DA, Harnett BM, Horswell R, Huang N, Johnson SG, Kahn M, Khanipov K, Kieler C, Luzuriaga KRD, Maidlow S, Martinez A, Mathew J, McClay JC, McMahan G, Melancon B, Meystre S, Miele L, Morizono H, Pablo R, Patel L, Phuong J, Popham DJ, Pulgarin C, Santos C, Sarkar IN, Sazo N, Setoguchi S, Soby S, Surampalli S, Suver C, Vangala UMR, Visweswaran S, Oehsen JV, Walters KM, Wiley L, Williams DA, Zai A. Synergies between centralized and federated approaches to data quality: a report from the national COVID cohort collaborative. Journal of the American Medical Informatics Association 2022 Mar 15;29(4):609–618. doi: 10.1093/jamia/ocab217

2. Prayitno, Shyu C-R, Putra KT, Chen H-C, Tsai Y-Y, Hossain KSMT, Jiang W, Shae Z-Y. A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications. Applied Sciences 2021 Nov 25;11(23):11191. doi: 10.3390/app112311191

3. Sebire NJ, Cake C, Morris AD. HDR UK supporting mobilising computable biomedical knowledge in the UK. BMJ Health Care Inform 2020 Jul;27(2):e100122. doi: 10.1136/bmjhci-2019-100122

4. Morrato EH, Lennox LA, Sendro ER, Schuster AL, Pincus HA, Humensky J, Firestein GS, Nadler LM, Toto R, Reis SE. Scale-up of the Accrual to Clinical Trials (ACT) network across the Clinical and Translational Science Award Consortium: a mixed-methods evaluation of the first 18 months. J Clin Trans Sci 2020 Dec;4(6):515–528. doi: 10.1017/cts.2020.505

5. Litton J-E. Launch of an Infrastructure for Health Research: BBMRI-ERIC. Biopreservation and Biobanking 2018 Jun;16(3):233–241. doi: 10.1089/bio.2018.0027

6. AKTIN and SPoCK Research Group, Bienzeisler J, Triefenbach L, Kombeiz A, Lottes M, Vogel C, Grabenhenrich L, Fischer M, Kocher T, Niekrenz L, Dreher M, Müller C, Röhrig R, Majeed RW. A Federated and Distributed Data Management Infrastructure to Enable Public Health Surveillance from Intensive Care Unit Data. In: Séroussi B, Weber P, Dhombres F, Grouin C, Liebe J-D, Pelayo S, Pinna A, Rance B, Sacchi L, Ugon A, Benis A, Gallos P, editors. Studies in Health Technology and Informatics IOS Press; 2022. doi: 10.3233/SHTI220507

7. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. Journal of the American Medical Informatics Association 2014 Jul 1;21(4):578–582. doi: 10.1136/amiajnl-2014-002747

8. Lawrence AK, Selter L, Frey U. SPHN - The Swiss Personalized Health Network Initiative. Stud Health Technol Inform 2020 Jun 16;270:1156–1160. PMID:32570562

9. Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med Schattauer GmbH; 2018 May;57(S 1):e50–e56. doi: 10.3414/ME18-03-0003

10. Gruendner J, Deppenwiese N, Folz M, Köhler T, Kroll B, Prokosch H-U, Rosenau L, Rühle M, Scheidl M-A, Schüttler C, Sedlmayr B, Twrdik A, Kiel A, Majeed RW. The Architecture of a

Feasibility Query Portal for Distributed COVID-19 Fast Healthcare Interoperability Resources (FHIR) Patient Data Repositories: Design and Implementation Study. JMIR Med Inform 2022 May 25;10(5):e36709. doi: 10.2196/36709

11. Benson T, Grieve G. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR. Cham: Springer International Publishing; 2016. doi: 10.1007/978-3-319-30370-3ISBN:978-3-319-30368-0

12. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, Welebob E, Scarnecchia T, Woodcock J. Advancing the Science for Active Surveillance: Rationale and Design for the Observational Medical Outcomes Partnership. Ann Intern Med 2010 Nov 2;153(9):600. doi: 10.7326/0003-4819-153-9-201011020-00010

13. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association 2010 Mar 1;17(2):124–130. doi: 10.1136/jamia.2009.000893

14. Kalra D, Beale T, Heard S. The openEHR Foundation. Stud Health Technol Inform 2005;115:153–173. PMID:16160223

15. Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, Kohane IS. The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories. Journal of the American Medical Informatics Association 2009 Sep 1;16(5):624–630. doi: 10.1197/jamia.M3191

16. Topaloglu U, Palchuk MB. Using a Federated Network of Real-World Data to Optimize Clinical Trials Operations. JCO Clinical Cancer Informatics 2018 Dec;(2):1–10. doi: 10.1200/CCI.17.00067

17. Scheufele E, Aronzon D, Coopersmith R, McDuffie MT, Kapoor M, Uhrich CA, Avitabile JE, Liu J, Housman D, Palchuk MB. tranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform. AMIA Jt Summits Transl Sci Proc 2014;2014:96–101. PMID:25717408

18. Lablans M, Kadioglu D, Mate S, Leb I, Prokosch H-U, Ückert F. Strategien zur Vernetzung von Biobanken: Klassifizierung verschiedener Ansätze zur Probensuche und Ausblick auf die Zukunft in der BBMRI-ERIC. Bundesgesundheitsbl 2016 Mar;59(3):373–378. doi: 10.1007/s00103-015-2299-y

19. Schüttler C, Prokosch H-U, Hummel M, Lablans M, Kroll B, Engels C, on behalf of the German Biobank Alliance IT development team. The journey to establishing an IT-infrastructure within the German Biobank Alliance. Ahmed Z, editor. PLoS ONE 2021 Sep 22;16(9):e0257632. doi: 10.1371/journal.pone.0257632

20. Schüttler C, Huth V, Von Jagwitz-Biegnitz M, Lablans M, Prokosch H-U, Griebel L. A Federated Online Search Tool for Biospecimens (Sample Locator): Usability Study. J Med Internet Res 2020 Aug 18;22(8):e17739. doi: 10.2196/17739

21. ATLAS. GitHub. Available from: https://github.com/OHDSI/Atlas/wiki/Home [accessed Jan 2, 2024]

22. Hoffmann J, Hanß S, Kraus M, Schaller J, Schäfer C, Stahl D, Anker SD, Anton G, Bahls T, Blankenberg S, Blumentritt A, Boldt L-H, Cordes S, Desch S, Doehner W, Dörr M, Edelmann F, Eitel I, Endres M, Engelhardt S, Erdmann J, Eulenburg K, Falk V, Felix SB, Frank D, Franke T, Frey N, Friede T, Geidel L, Germans L, Grabmaier U, Halle M, Hausleiter J, Jakobi V, Jebran A-F, Jobs A, Kääb S, Karakas M, Katus HA, Klatt A, Knosalla C, Krebser J, Landmesser U, Lee M, Lehnert K, Lesser S, Leyh K, Lorbeer R, Mach-Kolb S, Meder B, Nagel E, Nolte CH, Parwani AS, Petersmann A, Puls M, Rau H, Reiser M, Rienhoff O, Scharfe T, Schattschneider M, Scheel H, Schnabel RB, Schuster A, Schmitt B, Seidler T, Seiffert M, Stähli B-E, Stas A, J. Stocker T, Von Stülpnagel L, Thiele H, Wachter R, Wakili R, Weis T, Weitmann K, Wichmann H-E, Wild P, Zeller T, Hoffmann W, Zeisberg EM, Zimmermann W-H, Krefting D, Kühne T, Peters A, Hasenfuß G, Massberg S, Sommer T, Dimmeler S, Eschenhagen T, Nauck M. The DZHK research platform: maximisation of scientific value by enabling access to health data and biological samples collected in cardiovascular clinical studies. Clin Res Cardiol 2023 Jul;112(7):923–941. doi: 10.1007/s00392-023-02177-5

23. Home | ClinicalTrials.gov. Available from: https://clinicaltrials.gov/ [accessed Mar 11, 2024]

24. Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. Summit Transl Bioinform 2010 Mar 1;2010:46–50. PMID:21347148

25. Gulden C, Mate S, Prokosch H-U, Kraus S. Investigating the Capabilities of FHIR Search for Clinical Trial Phenotyping. German Medical Data Sciences: A Learning Healthcare System IOS Press; 2018;3–7. doi: 10.3233/978-1-61499-896-9-3

26. Schüttler C, Prokosch H-U, Sedlmayr M, Sedlmayr B. Evaluation of Three Feasibility Tools for Identifying Patient Data and Biospecimen Availability: Comparative Usability Study. JMIR Med Inform 2021 Jul 21;9(7):e25531. doi: 10.2196/25531

27. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, Duffett C, Dube K, Gallagher T, McLachlan S. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc Oxford Academic; 2018 Mar 1;25(3):230–238. doi: 10.1093/jamia/ocx079

28. Drenkhahn C, Ohlsen T, Wiedekopf J, Ingenerf J. WASP—A Web Application to Support Syntactically and Semantically Correct SNOMED CT Postcoordination. Applied Sciences 2023 May 16;13(10):6114. doi: 10.3390/app13106114

29. Release v1.0.0 · medizininformatik-initiative/clinical-cohort-definition-language. Available from: https://github.com/medizininformatik-initiative/clinical-cohort-definition-language/releases/tag/v1.0.0 [accessed Mar 18, 2024]

30. medizininformatik-initiative/feasibility-deploy. Medizininformatik-Initiative; 2024. Available from: https://github.com/medizininformatik-initiative/feasibility-deploy [accessed Jun 16, 2024]

31. medizininformatik-initiative/feasibility-gui. Medizininformatik-Initiative; 2024. Available from: https://github.com/medizininformatik-initiative/feasibility-gui [accessed Jun 16, 2024]

32. Sedlmayr B, Sedlmayr M, Kroll B, Prokosch H-U, Gruendner J, Schüttler C. Improving COVID-19 Research of University Hospitals in Germany: Formative Usability Evaluation of the CODEX Feasibility Portal. Appl Clin Inform 2022 Mar;13(02):400–409. doi: 10.1055/s-0042-1744549

33. Schüttler C, Zerlik M, Gruendner J, Köhler T, Rosenau L, Prokosch H-U, Sedlmayr B. Empowering Researchers to Query Medical Data and Biospecimens by Ensuring Appropriate Usability of a Feasibility Tool: Evaluation Study. JMIR Hum Factors 2023 Apr 19;10:e43782. doi: 10.2196/43782

34. Gebhardt M, Gruendner J, Kleinert P, Buckow K, Rosenau L, Semler SC. Towards a National Portal for Medical Research Data (FDPG): Vision, Status, and Lessons Learned. Studies in Health Technology and Informatics IOS Press; 2023. doi: 10.3233/SHTI230124

35. flare/.github/integration-test at main · medizininformatik-initiative/flare. GitHub. Available from: https://github.com/medizininformatik-initiative/flare/tree/main/.github/integration-test [accessed Jun 16, 2024]

36. Rosenau L, Majeed RW, Ingenerf J, Kiel A, Kroll B, Köhler T, Prokosch H-U, Gruendner J. Generation of a Fast Healthcare Interoperability Resources (FHIR)-based Ontology for Federated Feasibility Queries in the Context of COVID-19: Feasibility Study. JMIR Med Inform 2022 Apr 27;10(4):e35789. doi: 10.2196/35789

37. medizininformatik-initiative/flare: Feasibility Analysis Request Executor. Available from: https://github.com/medizininformatik-initiative/flare [accessed Jan 4, 2024]

38. medizininformatik-initiative/sq2cql. Available from: https://github.com/medizininformatik-initiative/sq2cql [accessed Jan 4, 2024]

39. Rosenau L, Ingenerf J. Structured Queries to AQL: Querying OpenEHR Data Leveraging a FHIR-Based Infrastructure for Federated Feasibility Queries. MEDINFO 2023 — The Future Is Accessible IOS Press; 2024. p. 33–37. doi: 10.3233/SHTI230922

40. Grimes J, Szul P, Metke-Jimenez A, Lawley M, Loi K. Pathling: analytics on FHIR. J Biomed Semant 2022 Sep 8;13(1):23. doi: 10.1186/s13326-022-00277-1

41. Rosenau L, Majeed RW, Ingenerf J, Kiel A, Kroll B, Köhler T, Prokosch H-U, Gruendner J. Generation of a Fast Healthcare Interoperability Resources (FHIR)-based Ontology for Federated Feasibility Queries in the Context of COVID-19: Feasibility Study. JMIR Med Inform 2022 Apr 27;10(4):e35789. doi: 10.2196/35789

42. 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

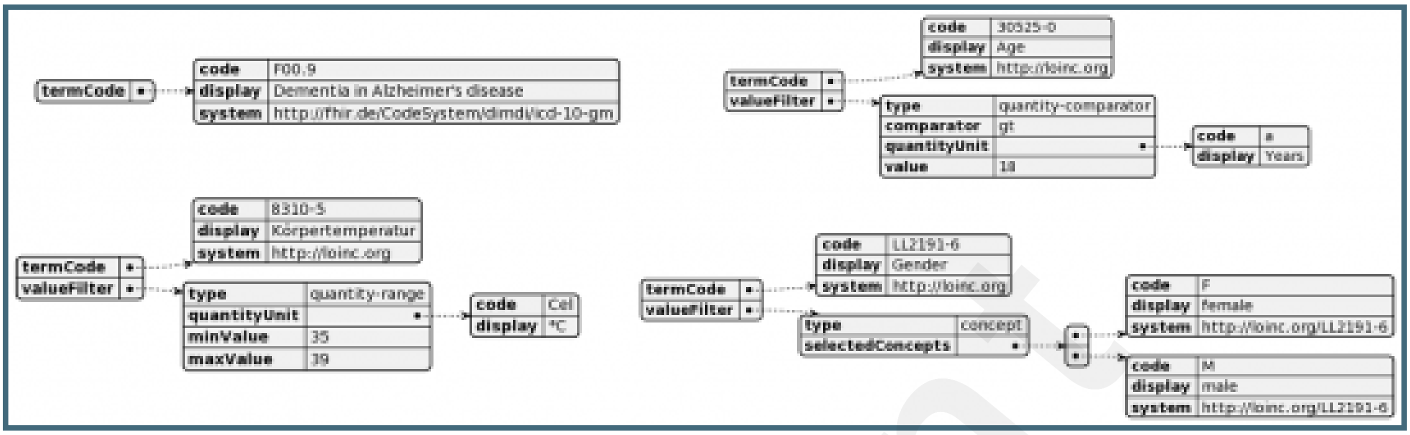
43. Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, Makadia R, Jin P, Shang N, Kang T, Weng C. Criteria2Query: a natural language interface to clinical databases for cohort definition. Journal of the American Medical Informatics Association 2019 Apr 1;26(4):294–305. doi: 10.1093/jamia/ocy178

44. Fang Y, Idnay B, Sun Y, Liu H, Chen Z, Marder K, Xu H, Schnall R, Weng C. Combining human and machine intelligence for clinical trial eligibility querying. Journal of the American Medical Informatics Association 2022 Jun 14;29(7):1161–1171. doi: 10.1093/jamia/ocac051

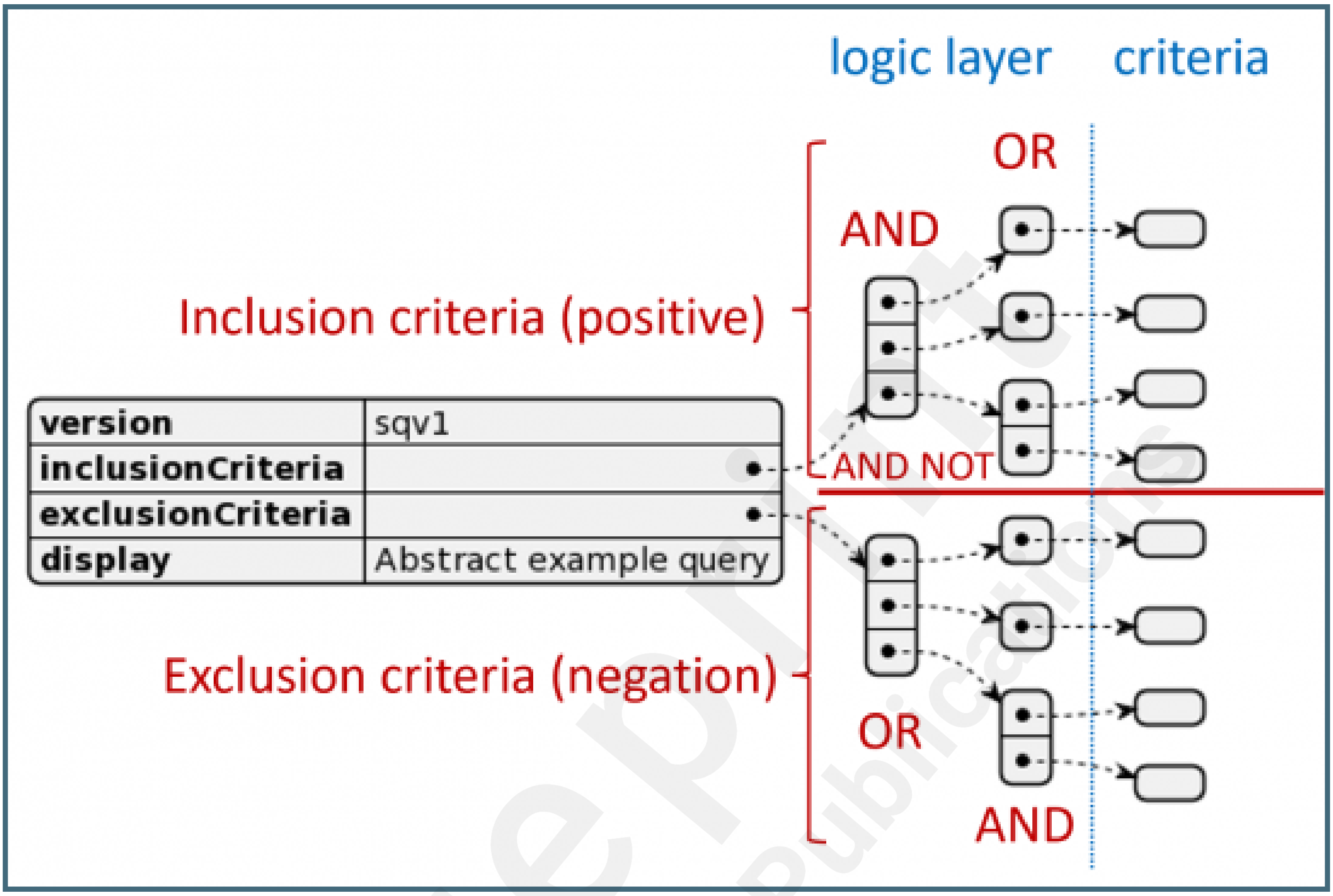
Supplementary Files

Figures

Different types of criteria definitions. A: Simple conceptual criterion. B: Numeric criterion with quantitative comparison. C: Numeric criterion with range restriction. D: Valueset criterion.

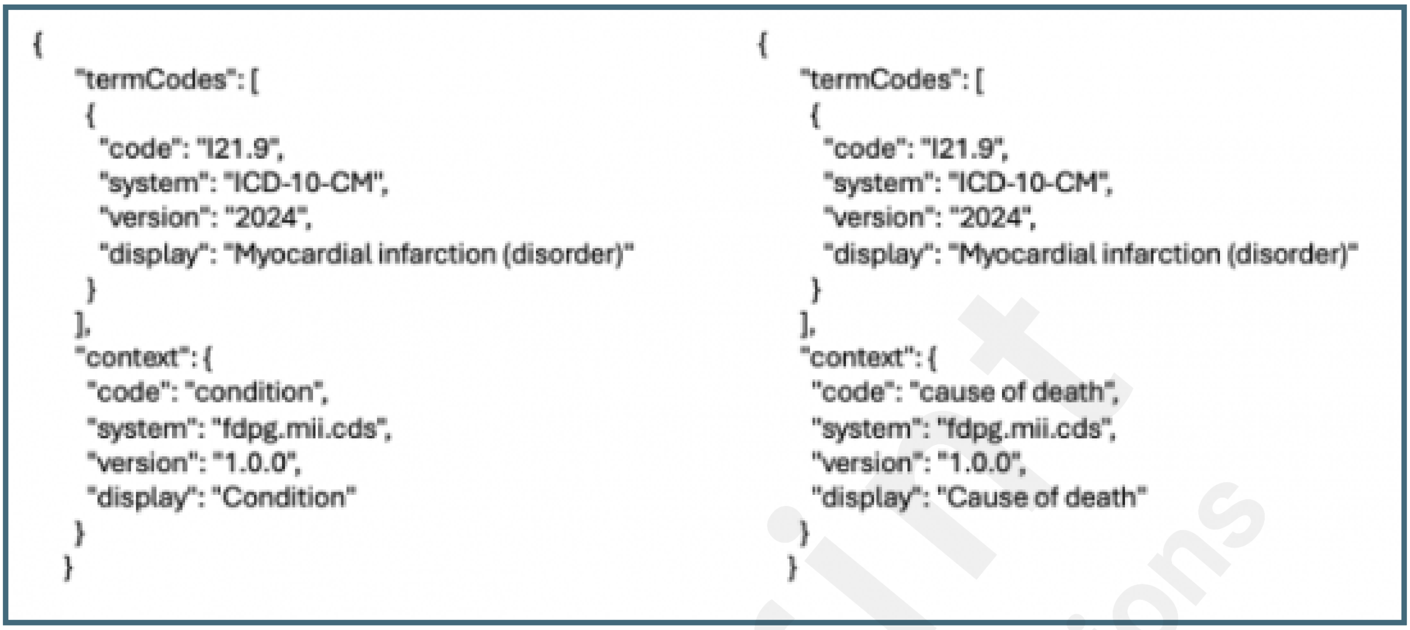


Structured query syntax top-level elements and logic layer. Certain criterion types will imply additional intrinsic logical relations. See Valueset-Criteria and attribute filters and time restrictions.



version	sqv1
inclusionCriteria	●
exclusionCriteria	●
display	Abstract example query

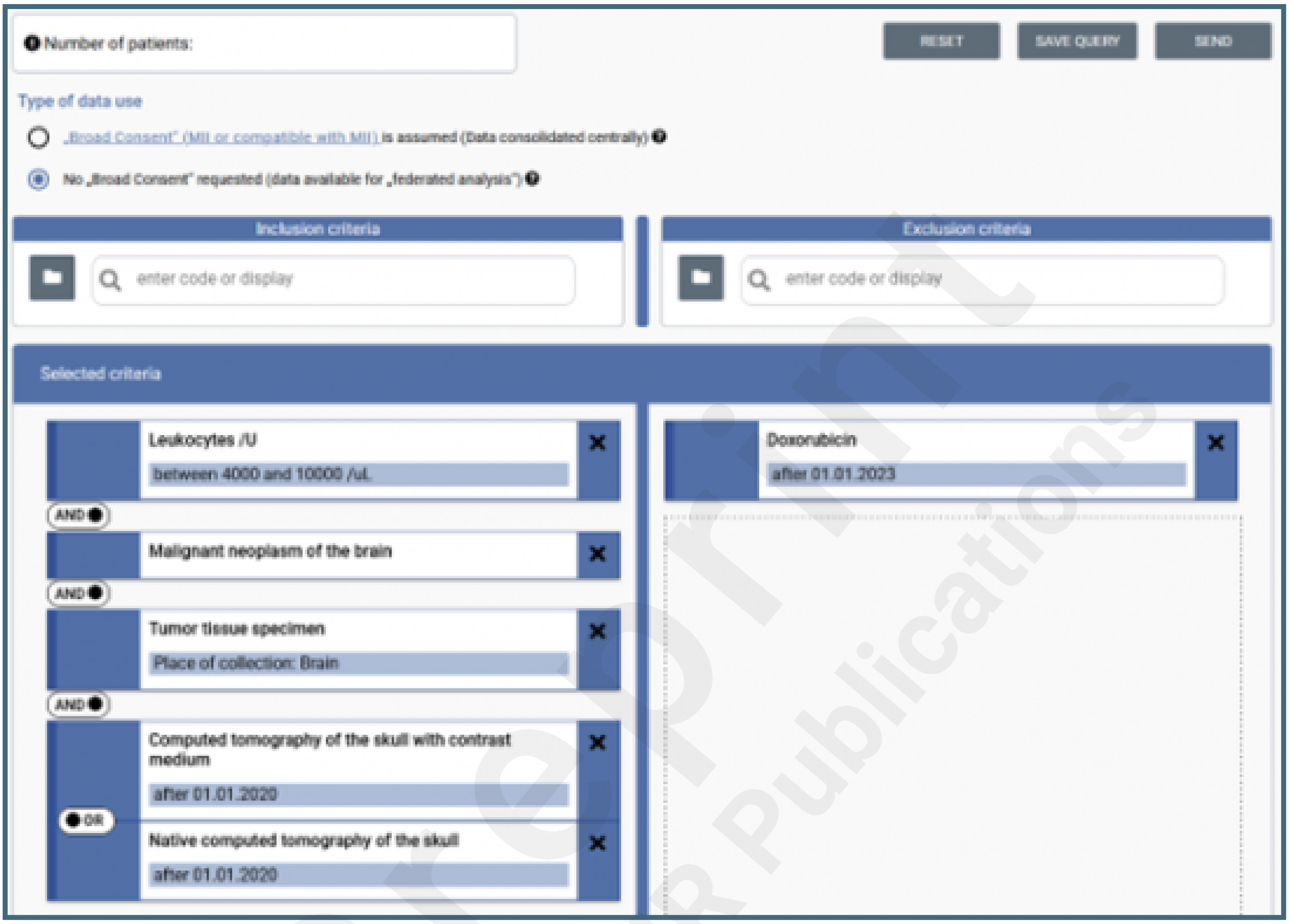
Myocardial infarction in two contexts (condition and cause of death).



```
{
    "termCodes": [
     {
      "code": "I21.9",
      "system": "ICD-10-CM",
      "version": "2024",
      "display": "Myocardial infarction (disorder)"
     }
    ],
    "context": {
     "code": "condition",
     "system": "fdpg.mii.cds",
     "version": "1.0.0",
     "display": "Condition"
    }
   }
```

```
{
    "termCodes": [
     {
      "code": "I21.9",
      "system": "ICD-10-CM",
      "version": "2024",
      "display": "Myocardial infarction (disorder)"
     }
    ],
    "context": {
     "code": "cause of death",
     "system": "fdpg.mii.cds",
     "version": "1.0.0",
     "display": "Cause of death"
    }
   }
```

Example of a feasibility query in the FDPG Feasibility Portal to find patients with a leukocyte count within a normal range, with a malignant neoplasm of the brain, available tumor tissue specimen, and a CT scan after 01.01.2020 who did not take Doxorubicin.



Number of patients:

Type of data use

○ „Broad Consent" (MII or compatible with MII) is assumed (Data consolidated centrally)

● No „Broad Consent" requested (data available for „federated analysis")

Leukocytes /U
between 4000 and 10000 /uL

AND

Malignant neoplasm of the brain

AND

Tumor tissue specimen
Place of collection: Brain

AND

Computed tomography of the skull with contrast medium
after 01.01.2020

OR

Native computed tomography of the skull
after 01.01.2020

Doxorubicin
after 01.01.2023

Multimedia Appendixes

Example of specimen criteria with an ICD-o-3 attribute indicating the location the specimen was taken from.
URL: http://asset.jmir.pub/assets/cd9c0a2f9349fd1fe0c63eff6d162566.png

TOC/Feature image for homepages

Untitled.