applied sciences

MDPI

Article

Towards an Ontology-Based Phenotypic Query Model

Christoph Beger [1,2,3,*], Franz Matthies [1,3], Ralph Schäfermeier [1,3], Toralf Kirsten [1,3,4], Heinrich Herre [1] and Alexandr Uciteli [1,3,*]

1 Institute for Medical Informatics, Statistics and Epidemiology (IMISE), Leipzig University, 04107 Leipzig, Germany; franz.matthies@imise.uni-leipzig.de (F.M.); ralph.schaefermeier@uni-leipzig.de (R.S.); toralf.kirsten@imise.uni-leipzig.de (T.K.); heinrich.herre@imise.uni-leipzig.de (H.H.)
2 Growth Network CrescNet, Leipzig University, 04103 Leipzig, Germany
3 SMITH Consortium of the German Medical Informatics Initiative, 04103 Leipzig, Germany
4 Department Medical Data Science, Leipzig University Medical Center, 04107 Leipzig, Germany
* Correspondence: christoph.beger@imise.uni-leipzig.de (C.B.); auciteli@imise.uni-leipzig.de (A.U.)

Abstract: Clinical research based on data from patient or study data management systems plays an important role in transferring basic findings into the daily practices of physicians. To support study recruitment, diagnostic processes, and risk factor evaluation, search queries for such management systems can be used. Typically, the query syntax as well as the underlying data structure vary greatly between different data management systems. This makes it difficult for domain experts (e.g., clinicians) to build and execute search queries. In this work, the Core Ontology of Phenotypes is used as a general model for phenotypic knowledge. This knowledge is required to create search queries that determine and classify individuals (e.g., patients or study participants) whose morphology, function, behaviour, or biochemical and physiological properties meet specific phenotype classes. A specific model describing a set of particular phenotype classes is called a Phenotype Specification Ontology. Such an ontology can be automatically converted to search queries on data management systems. The methods described have already been used successfully in several projects. Using ontologies to model phenotypic knowledge on patient or study data management systems is a viable approach. It allows clinicians to model from a domain perspective without knowing the actual data structure or query language.

Keywords: eligibility determination; biomedical ontologies; electronic health records; health information interoperability; information storage and retrieval; health level seven; FAIR data principles

Citation: Beger, C.; Matthies, F.; Schäfermeier, R.; Kirsten, T.; Herre, H.; Uciteli, A. Towards an Ontology-Based Phenotypic Query Model. *Appl. Sci.* **2022**, *12*, 5214. https://doi.org/10.3390/app12105214

Academic Editor: Enno van der Velde

Received: 13 April 2022
Accepted: 16 May 2022
Published: 21 May 2022

1. Introduction

Clinical research based on data from clinical trial data management systems and hospital information systems (HIS) plays an important role in transferring basic findings into the daily practices of physicians. To perform clinical trials, recruitment of patients is a crucial task, and if the recruitment is insufficient, this can lead to delayed or even failed clinical studies [1,2]. The increasing introduction of electronic health record (EHR) systems in HIS and proliferation of clinical data can enhance and accelerate cohort identifications [3,4] and enable the implementation of electronic screenings, which can improve recruitment efficiency [5]. In addition to trial recruitment, electronic screenings can further be used for the diagnostic process and the evaluation of risk factors (i.e., finding patients with combinations of clinical features that can indicate a harmful vital state).

Electronic screenings can be expressed with search queries that are executable on any type of trial or clinical data management system. Each system has its own query language, but some systems can share a common language. Queries are typically searching for single clinical features or combinations and can specify the including and excluding criteria (i.e., eligibility criteria). It usually takes a lot of effort for researchers to build search queries for

eligibility criteria [3]. For instance, descriptive metadata of clinical features may be missing, or it may be unclear how to reference features in queries. Further information needs to be included in search queries, such as units of measurement, patient age restrictions, or codes of external terminologies. Queries may consist of complex logical combinations (conjunctions and disjunctions) of clinical features. We consider such features in the context of standardised ontology-based phenotype models that serve as representations of the phenotypic knowledge required for search queries. According to the definition of Scheuermann et al., a phenotype is a "(combination of) bodily feature(s) of an organism determined by the interaction of its genetic make-up and environment" [6] and can be interpreted as clinical features. Therefore, phenotypes can be used as filter criteria for search queries. Definitions of phenotypes need to be sharable and of a high quality to be properly usable for the characterisation of patient cohorts in large electronic health record repositories [7,8] and to fulfil the promise of leveraging EHR data into important individual and population health [9].

The formal representation of eligibility criteria is a topic of current research, and several groups are developing formalisms for knowledge representation [10,11]. Several existing languages are promising candidates for expressing eligibility criteria, such as SPARQL [12], Arden syntax [13], Structured Query Language (SQL), and Description Logic. In particular, SPARQL is a state-of-the-art approach in the Semantic Web [14] community for defining queries on triple data, and it can also be used for the Ontology-Based Data Access (OBDA) [15,16] paradigm.

In this work, we propose an ontology-based approach that enables researchers to build computable representations of phenotypes [17]. These representations can be transformed into the query languages of trial or clinical data management systems, with ontologies enabling class-based reasoning and annotation with rich metadata. Therefore, the focus is not on the form of the patient data (e.g., databases, triple stores, FHIR repositories, or file formats) or the used query languages (SPARQL, SQL, etc.). It is rather on how complex phenotypic knowledge can be constructed, shared, and semantically represented in a way comprehensible to domain experts and reusable for query building.

2. Materials and Methods

2.1. Approach Overview

Phenotypes are the observable characteristics [18–20] of an individual. These include clinical conditions [6] as they have been obtained from clinical patient care processes. Therefore, we consider "phenotypic queries" as queries that can be used to search for individuals (patients or study participants) with specific phenotypes. A scientist could be interested, for example, in finding male individuals 20–40 years old and having a BMI greater than 30 kg/m^2 and diagnosed heart failure.

By building phenotypic queries in a machine-readable format, they become executable on electronically available data (e.g., electronic health records). Phenotypic queries return a set of individuals meeting the criteria described by the query. Thus, the goal of such phenotypic queries is to return a subset of the available individual's data mostly represented by a corresponding (patient) identifier. In a further step, this identifier set can be used to retrieve additional data for the selected individuals.

Phenotypic queries can be used in the research context, in patient care, and in administrative processes. For example, cohorts for studies are definable [21], which may consist of complex concatenations of phenotype classes. On the other hand, treatment centres could implement queries to stratify patients into phenotype (sub)classes and thus provide the best available care for each of them [22]. Finally, phenotypic queries may be used as a data curation tool. Consider a scenario where (combinations of) implausible value ranges are specified in phenotypic queries that are used to search for corresponding (possibly incorrect) patient data in EHRs.

We propose a multi-level approach allowing the specification of phenotypic knowledge on an abstract level first. The knowledge is then transformed into specific, already

available, and often widely used query languages such as SPARQL for RDF- or OWL-based data, SQL for Relational Database Management Systems (RDMS), and HL7 FHIR Search for FHIR systems. Queries in these query languages can then be directly executed within specific data management systems. The advantage of this approach is twofold. First, phenotypic knowledge can be formally specified without any implementation-specific syntax necessary for the used data management system. By separating the formal aspect from implementation specialities, the end users (especially non-computer scientists) can model such knowledge without knowing the implementation syntax. By providing translator and adapter tools, the phenotypic knowledge can then be converted into desired query languages (e.g., SPARQL, SQL, or FHIR Search). Second, such ontological specifications and their parts (e.g., single classes) are reusable and applicable on different source systems by using suitable adapters.

Following this multi-level approach (see Figure 1 for an overview), we use the Core Ontology of Phenotypes (COP) [23] on the top level as a general phenotypic model. The COP provides the basic building blocks (i.e., types of classes, properties, and axioms) and a common structure (i.e., taxonomy and other relations) which must be used by each use case's specific phenotypic model. Each specific model is in turn managed in a Phenotype Specification Ontology (PheSO) that integrates the COP and follows the structure the COP provides (e.g., the PheSO classes are subclasses of the COP classes). The translator tools are based on the COP (i.e., they know the general structure of all PheSOs) and can generate queries from PheSOs in a corresponding query language.

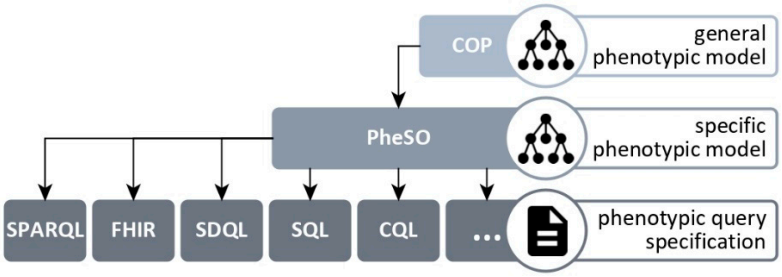

Figure 1. Overview of the components used to build phenotypic queries. The Core Ontology of Phenotypes (COP) serves as a general phenotypic model and is used by Phenotype Specification Ontologies (PheSO) to build use case-specific models that can be converted to queries.

We describe the COP and the query building process in more detail throughout the next subsections.

2.2. Core Ontology of Phenotypes

The Core Ontology of Phenotypes (COP) [23] is an ontological framework to model different types of phenotype classes. It serves as a general model to express phenotypic knowledge.

In the COP, phenotypes are defined as dependent individuals, such as the body temperature of a person. Those individuals are instances of phenotype classes. In our example, the phenotype class "body temperature" possesses individual temperatures of people as instances. We distinguish between single, combined, and derived phenotype classes. Single phenotypes are the single or atomic properties of an organism. The body temperature, height, and weight of a person are single phenotypes. Some single phenotypes must be considered together in a specific context (e.g., for reasons of joint data acquisition or further analyses). The combination of the two single phenotypes "gender" and "waist circumference", for example, can imply a substantially increased risk of metabolic complications if the gender is male and the waist circumference is greater than 102 cm [24]. Hence, it is valuable to manage and ontologically model such combinations. They are represented as combined phenotype classes in the COP. Derived phenotype classes represent the additional derived properties of an organism (e.g., BMI). The derivation rule is specified as

a mathematical formula taking single phenotype classes as variables (e.g., BMI = weight (kg)/height $(m)^2$).

These phenotype classes are unrestricted (i.e., they define properties for which values can be obtained, either originally measured or derived). For example, instances of the single phenotype class "body temperature" are the actual temperature values of all living beings without any restriction. Subclasses of these phenotype classes may partition the feature space by defining specific conditions, such as "high body temperature" for a body temperature greater than 37.5 °C. Therefore, these subclasses are referred to as restricted phenotype classes. They are available for all three phenotype class types and thus named accordingly as restricted single, restricted combined, and restricted derived phenotype classes. The mentioned restrictions may be range restrictions or discrete value sets.

Phenotype classes possess various common attributes such as labels, descriptions, and codes of external concepts. Other attributes vary depending on the type of the phenotype class; some of them are mandatory for the later translation into the queries specific to the available query languages. To these phenotype class-specific attributes (aside from the identifier, name, etc.) belong the following:

- Unrestricted single phenotype (USiP) class: a data type, optionally a unit of measure, and an aggregate function. For example, the phenotype class "body temperature" having the data type "numeric", unit "°C", and aggregate function "average" to handle multiple values returned by a query.
- Restricted single (RSiP) and derived phenotype (RDeP) class: a restriction to partition the feature space of the corresponding USiP or UDeP, such as a value greater than 37.5 for the restricted phenotype class "high body temperature".
- Restricted combined phenotype (RCoP) class: a Boolean expression, such as "sex is male" or "high body temperature".
- Unrestricted derived phenotype (UDeP) class: a mathematical formula and Boolean expression consisting of AND-linked variables used in the formula (e.g., the formula "weight (kg) / height $(m)^2$" and the variables "weight" and "height" for the phenotype class "BMI".

The simple attributes of the phenotype classes are defined as annotations. The logical relations between the phenotype classes as well as range restrictions are represented in OWL by anonymous equivalent classes or general class axioms based on property restrictions.

Let us consider an example study for people with hypertension or obesity as inclusion criteria. All relevant phenotype classes are modelled in a single PheSO based on the COP. First, the USiP classes "Age", "Gender", "Weight", "Height", "Systolic_Blood_Pressure", "Myocardial_Infarction", and "Stroke" are added (Figure 2A), including relevant annotations such as data types, units, and codes. Some example annotations of the class "Systolic_Blood_Pressure" are listed in Figure 2A1. Assuming men between 40 and 65 years (i.e., age in years $\geq$40 and <66) with high blood pressure are relevant for our example study, we insert the RSiP classes "Male", "Age_40_66", and "Hypertension" under the corresponding USiP classes (Figure 2A) and define the desired value ranges as property restrictions (Figure 2A2). Next, in order to define a phenotype class for obesity based on BMI, we specify the UDeP class "BMI" (Figure 2B) and annotate it with an appropriate formula (Figure 2B1). The USiP classes "Height" and "Weight" are used as variables in this formula. The RDeP class "Obesity" can now be added as a subclass of "BMI" with a respective value range (Figure 2B2). Finally, the UCoP class "SBP_BMI" ("SPB" stands for "systolic blood pressure") and its restricted subclass "Hypertension_OR_Obesity" are added (Figure 2C). The corresponding Boolean restriction is depicted in Figure 2C1.

For an in-depth description of the COP as well as further examples and details on how PheSOs are composed and what information must be provided by domain experts, please refer to [23].

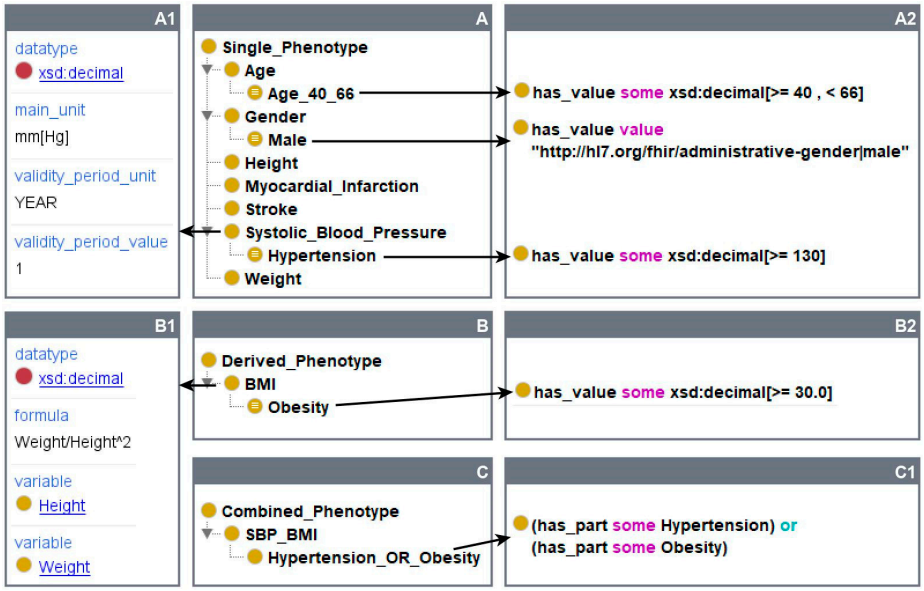

Figure 2. An example PheSO (some excerpts from the OWL representation and Protégé [25] screenshots). Subfigures (**A**–**C**) show all single, derived, and combined phenotype classes in the PheSO, respectively. (**A1**,**B1**) show annotations of the classes "Systolic_Blood_Pressure" and "BMI". (**A2**,**B2**,**C1**) show Manchester Syntax expressions for selected phenotype classes. The full ontology is available as Supplementary Material File S1 and at https://health-atlas.de/lha/8ATHQGDE4F-4 (accessed on 3 May 2022).

2.3. Classification of Phenotypic Queries

In this subsection, we propose a basic classification of phenotypic queries based on different kinds of phenotype classes. The knowledge required for the described query types can be modelled in Phenotype Specification Ontologies (PheSOs) according to the general phenotyping model provided by the COP. The specific model (PheSO) can then be used to generate query specifications in a multitude of syntaxes (e.g., SPARQL, SQL, or HL7 FHIR Search; see section applications).

As stated before, phenotypes are the observable bodily feature(s) of organisms (e.g., people, patients, or study participants). This includes morphology, function, behaviour, or biochemical and physiological properties [18–20]. According to the COP [23], phenotypes are representable as ontological classes. One can use these classes to build phenotypic queries, which determine and classify individuals (e.g., patients) whose properties meet specific phenotype classes. For example, the query "select males aged 12–18 with BMI greater than 30 kg/m^2" can be expressed with the phenotype classes sex, age, height, weight, and body mass index.

We distinguish between several types of phenotypic queries, which are described below. The relations of the query types to the phenotype classes are depicted in Figure 3.

2.3.1. Single Phenotype Queries

An **unrestricted single phenotype (USiP) query** represents a search for people that possess a certain characteristic (e.g., diabetes diagnosis, a certain laboratory value, or a particular medication). Additionally, USiP queries can narrow the search using a period of time during which the characteristic must have been observed or recorded (e.g., diabetes diagnosis in the last year).

A **restricted single phenotype (RSiP) query** represents a search for people possessing a certain characteristic that have a value within a defined value range (e.g., the weight value must be between 70 and 80 kg, or a laboratory value must lie within a defined normal range). RSiP queries can also integrate a time restriction (e.g., the weight value must be between 70 and 80 kg in the last year). Additionally, it can be distinguished between

all-quantor and existence-quantor (i.e., whether all returned values or at least one value of a given characteristic must lie within a defined value range, such as all temperature observations or at least one temperature observation on the last day needing to be between 37.5 and 39 °C).

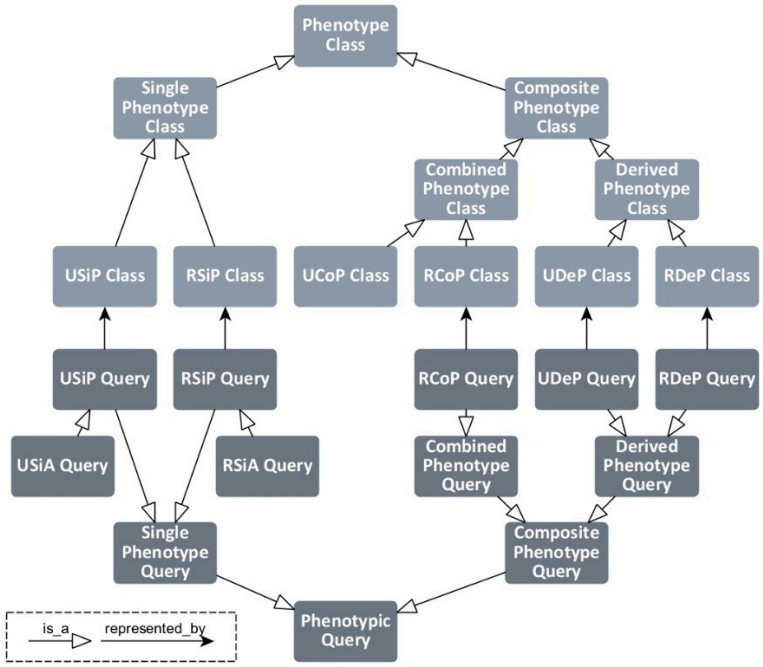

Figure 3. Displayed are types of phenotype classes and phenotypic queries as well as their relationships. Abbreviations used here are further described in upcoming sections.

An **unrestricted single aggregate (USiA) query** is a subtype of USiP query and defines an additional aggregate function. The aggregate value will be calculated from multiple values of the given patient characteristic and will be returned instead of all values of the characteristic (e.g., average, min, max, first, last, or count temperature). A time restriction is also possible (e.g., the average temperature on the last day).

A **restricted single aggregate (RSiA) query** is a subtype of the RSiP query. The aggregate value calculated from multiple values of the given patient characteristic must lie within a defined value range (e.g., the average temperature must be between 37.5 and 39 °C or with a time restriction, such as the average temperature on the last day needing to be between 37.5 and 39 °C).

2.3.2. Combined Phenotype Queries

A **restricted combined phenotype (RCoP) query** defines a Boolean expression (e.g., age > 70 years AND overweight). The parts of the expression represent single phenotype queries. To find people meeting all the criteria of an RCoP query, the corresponding single phenotype queries must be executed first. In our example, we would receive two result sets: one with people older than 70 years and one with overweight people. The AND operator between the single criteria is interpreted as an intersection operation on the corresponding sets. The OR operator would result in a set–theoretical union.

2.3.3. Derived Phenotype Queries

An **unrestricted derived phenotype (UDeP) query** uses a mathematical formula (e.g., BMI = weight (kg)/height $(m)^2$) to calculate or derive and return a new value (derivative). First, the corresponding USiP queries for the variables of the formula ("weight" and "height") must be executed (i.e., find most recent observations of weight and height for each patient). Then, the derived value (BMI) is calculated for each person and returned.

A **restricted derived phenotype (RDeP) query** defines a value range in which the calculated value must lie (e.g., BMI > 25). After executing the corresponding USiP queries for the variables and calculating the derivative (BMI) for each person, the initial person set (all people with documented weights and heights) is reduced to the people with a BMI > 25.

2.4. Inclusion and Exclusion Criteria

To search for a set of people based on multiple filter conditions, the individual phenotype classes described in Section 2.2. can serve as definitions of eligibility criteria (inclusion and exclusion). The classes can be combined to a superordinate specification of a complex phenotype class.

Let us consider the following eligibility criteria of an example study:

Inclusion criteria:

(ic1) male subjects
(ic2) age between 40 and 65 years (i.e., age in years $\geq$ 40 and <66)
(ic3) BMI $\geq$ 30 kg/m^2 (ic3a) OR at least one observation of systolic blood pressure $\geq$ 130 mmHg in the last month (ic3b)

Exclusion criteria:

(ec1) myocardial infarction
(ec2) stroke

The criteria (ec1) and (ec2) can be modelled using USiP classes because they just refer to the existence of an instance of a specific phenotype class (e.g., "stroke") without further restrictions. The gender (ic1) and age (ic2) restrictions are represented by RSiP classes. The Boolean expression (ic3) is covered by an RCoP class, whereby the BMI restriction (ic3a) is modelled using an RDeP class, and the blood pressure restriction (ic3b) is modelled using an RSiP class. The PheSO described in Section 2.2. and in Figure 2 can be used as a specific phenotypic model to cover the eligibility criteria introduced above. For this purpose, the classes "Gender", "Age", and "SPB_BMI" must only be annotated as inclusion criteria, and the classes "Myocardial_Infarction" and "Stroke" as exclusion criteria.

We outline an algorithm to search for eligible patients based on inclusion and exclusion criteria. 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). As a first step, we go through the list of inclusion criteria, generate and execute the corresponding queries, and determine the set–theoretical intersection of the received result sets. We call the result of the intersection the inclusion criteria result set (ICRS). Then, the single result sets of the exclusion criteria queries are removed (set–theoretical difference) from the ICRS. The remaining patient set fulfils all the inclusion and exclusion criteria.

The main result of this algorithm is a list of eligible patients. In a further step, the desired data of these patients can be queried. For this purpose, corresponding phenotype classes (i.e., USiP or UDeP classes) can be marked as "in projection". The term "projection" here refers to the projection operation in relational algebra or in database queries, which restricts the retrieved result set to certain attributes (in our case, to instances of desired phenotype classes). For the projected phenotype classes, the unrestricted queries (i.e., USiP, USiA, or UDeP queries) supplemented by the corresponding patient IDs are then generated and performed (e.g., search for weight values identified by LOINC code 3141-9 of the patients with IDs 1, 2, and 3).

2.5. From Idea to Result Set

This section outlines the steps and expected effort required by a facility holding patient data to define phenotypic knowledge, generate and execute the queries, and obtain a resulting data set. Each step has individual requirements for the expertise of the person performing it. The premise here is that a facility stores its patient data in a database system or software component with an application programming interface (API). APIs are software

interfaces that can be used by other software components to access functionalities. We will use them to search and access the facility's data.

First, there must be an interface in place which is capable of converting a PheSO into appropriate queries that are further on executed on the underlying patient data store. This interface must be implemented by a software developer. One can implement different translators or adapters from the ontological model to the respective query languages or just one translator (e.g., for SPARQL) but with different mappings of the respective data sources to a suitable representation (e.g., using OBDA [15,16] or direct transformation in RDF or OWL). This makes no difference in our approach. For instance, the developer can use the Phenotype Ontology Manager Core API (PhenoMan) [23], which is implemented in Java using the OWL API. PhenoMan was initially developed in the SMITH project [26], and it provides simple methods to read, write, and reason with phenotype classes in PheSOs. PhenoMan can be used to access the phenotype classes in a PheSO and convert them into API calls to the facility data store (e.g., FHIR Search queries). In case the data store supports standardised interfaces such as FHIR Search, such an interface may already exist and will only require a few customisations. An example of an FHIR Search connection is described in the results section.

Before queries can be generated and executed, a PheSO must be modelled by a domain expert. Depending on the patient cohort to be described by the query, additional literature research may be required to carefully collect phenotypes from previous scientific findings. By using tools such as Protégé [25] or the Phenotype Editor [23], and with the help of ontologists, domain experts create several single phenotype classes that serve as basic components for modelling composite phenotype classes. By marking specific classes as inclusion or exclusion criteria, a coherent description of the desired cohort is produced.

The amount of time required to build a PheSO highly depends on the complexity of the phenotype classes describing the cohort and varies from some minutes to a few hours. Formalised algorithms already available (e.g., Type 2 Diabetes Mellitus in PheKB [27]) can significantly simplify the modelling process, and previously built PheSOs may be reusable for multiple queries. Domain experts could also prepare a set of frequently used phenotype classes to streamline and speed up the modelling process.

Ideally, the patient data store interface has a graphical user interface through which domain experts can upload—or even build—their PheSO and execute it. The PheSO is then converted to one or multiple internal queries (i.e., API calls), and the resulting patient cohort is returned to the user. The user does not need to know the underlying query language used by the data store. The execution time depends on the type of the data store as well as the implementation of the interface and complexity of the query. Figure 4 shows the layout of a potential architecture for building PheSOs in a graphical user interface, transforming them into data source queries and retrieving the resulting data sets.

2.6. Towards FAIR Data Principles

The underlying representations of specific phenotypic knowledge are PheSOs. We want to enable and encourage scientists to publish and reuse these PheSOs so that the emergence of communities around ontological phenotyping is strengthened. To achieve this goal, we are developing our toolset according to the FAIR guiding principles for scientific data management [28,29]. These principles set fundamental requirements for scientific (meta)data and the sharing processes which, according to the letters of FAIR, comprise findability, accessibility, interoperability, and reusability, as proposed and described by Wilkinson et al. [28].

A side effect of FAIRifying phenotypic queries is that we are also FAIRifying sets of phenotypic models. This results from the fact that a PheSO contains phenotype classes which, per definition, represent all individual phenotypes. One can use these abstract sets and perform corresponding queries on actual data sources, such as an FHIR server (see the Results section). The result may be a set of patients or probands with phenotypes that are instances of the phenotype classes.

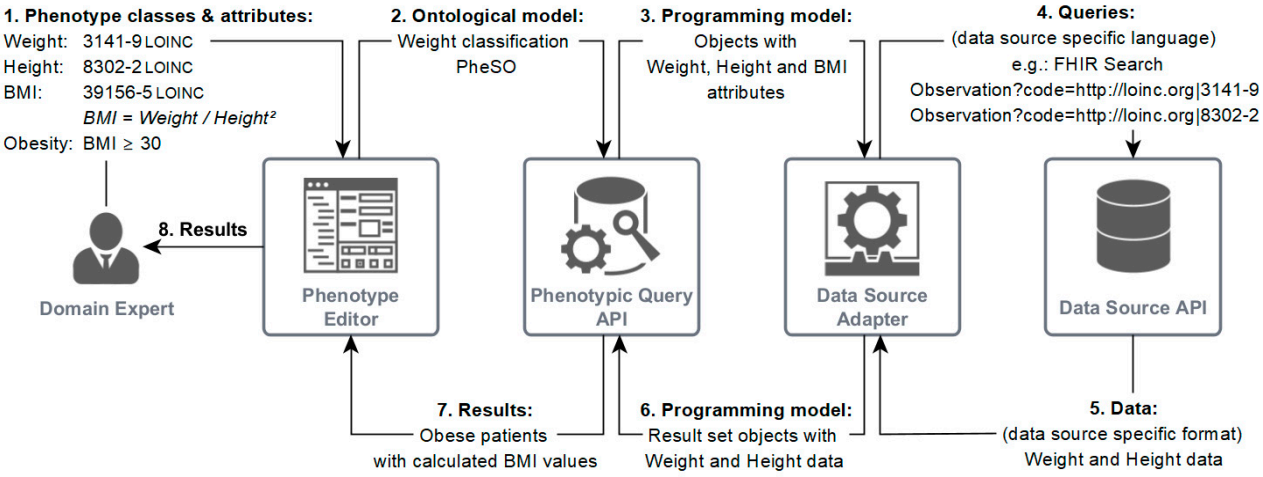

Figure 4. Possible architecture to implement the transformation of ontology-based phenotypic models to query languages and to obtain the resulting data sets. The data source adapter can be replaced by any kind of query-specific adapter (e.g., FHIR Search adapter).

In the following sections, we will outline which requirements of FAIR [28] are already met by our phenotype framework. Please keep in mind that the work described in this publication is only the first step to an extensive framework, and further work is needed to provide an ontology-based ecosystem for modelling, sharing, reusing, and executing phenotype algorithms and phenotypic queries.

2.6.1. Findable and Interoperable

Phenotype classes, as well as complete Phenotype Specification Ontologies, are assigned unique identifiers. This is accomplished via an ontology IRI for a PheSO and concatenation of the IRI and class name for a phenotype class. Those identifiers are currently chosen by the ontologist modelling the PheSO, and there is no system in place for checking the uniqueness. Thus, identifiers are not globally unique or persistent. Phenotypes are described with rich metadata by using terms of established terminologies such as DCMI Metadata [30] and the PROV Ontology [31], as well as medical terminologies (e.g., LOINC or SNOMED CT). Ontologies can be exported and stored in OWL 2 document format [32] or RDF. Both formats are used for knowledge representation and are popular solutions for storing FAIR (meta)data [33].

2.6.2. Accessible

Web services supporting the HTTP(S) protocol may be implemented, which respond to IRIs, being used as identifiers for phenotype classes and ontologies and the responding metadata of phenotype classes. This would make classes and ontologies retrievable. HTTP also fulfils the requirement of support for authentication and authorisation procedures.

2.6.3. Reusable

The community standard OWL2 is a sufficient format to facilitate reusability. We encourage domain experts and ontologists to include as many describing metadata attributes as possible when modelling phenotype classes. In particular, the annotation property "code" for referencing controlled vocabularies ensures that phenotype classes are sufficiently explained and interoperable with other data and software systems. Future web services providing access to and execution of PheSOs must store authorship details and allow users to specify licences; otherwise, reuse is not permitted. To enable the reuse of PheSOs as well as single phenotype classes, they should be shared in public repositories.

3. Results

In this section, we will outline some applications of the presented approach to model phenotypic knowledge and to transform it into the respective query languages of different data storage systems. For each specific query language (FHIR Search, SDQL, and SQL), we will present one example project where our phenotypic query approach has been used successfully.

3.1. Applications of the Phenotypic Query Approach

3.1.1. Use of HL7 FHIR Search in Smart Medical Information Technology for Healthcare to Enable Location-Independent Searches in Hospital Information Systems

The main occupation of university hospitals in Germany is the treatment of patients, but they are also involved in clinical research, and thus they must be able to provide data to clinical researchers. A local hospital information system typically consists of various specialised applications which result in very different IT architectures for each healthcare provider. The applications usually do not support a common transfer or use of data in electronic medical records [34,35].

To overcome difficulties (like heterogeneous data and structures) in using data from multiple hospital information systems for clinical research, the project Smart Medical Information Technology for Healthcare (SMITH) [26] was funded by the German Federal Ministry of Education and Research as part of the German Medical Information Initiative [36,37]. The objective of the SMITH consortium is to establish so-called Data Integration Centers (DICs) at all participating university hospitals, which will serve as common data access units. The DIC will maintain Health Data Storage (HDS), where a common core data set [38] of patient data will be stored and provided as standardised and interoperable EHRs. The EHRs are made accessible via the HL7 Fast Healthcare Interoperability Resources (FHIR) standard [39]. This is accomplished by incorporating an FHIR server into the local set-up. By using HL7 FHIR, SMITH will enable researchers to access available healthcare data in a standardised manner and use them for innovative research and treatment optimisations.

In SMITH, FHIR Search [40] is used to query the HDS. Searches are performed by sending GET requests to a RESTful FHIR server (like HAPI FHIR [41], VONK [42], or Blaze [43]). Those requests contain query parameters which specify the inclusion and exclusion criteria. Values like blood pressure can be encoded with codes of external terminologies like Logical Observation Identifiers Names and Codes (LOINC) and embedded into requests (e.g., 8480-6 for "blood pressure" in LOINC). The following example request would result in a response with all observations of blood pressure, where the measured values were greater or equal to 130 mmHg. Matching observations are returned in a so-called "bundle", a set of FHIR-formatted resources.

GET [base]/Observation?component-code-value-quantity=
http://loinc.org\T1\textbar{}8480-6$ge130\T1\textbar{}
http://unitsofmeasure.org\T1\textbar{}mm[Hg]

The mapping of PheSO phenotype classes to phenotypic queries and FHIR Search queries was achieved with the software suite PhenoMan [23,44]. To specify the FHIR-related conditions for the desired result set, domain experts can add additional properties to the phenotype classes in the PheSOs. The mandatory property "code" must be specified to map phenotypes to one or more external terminologies (e.g., LOINC). These codes are used to query specific observation types (e.g., blood pressure observation) and medical conditions (e.g., myocardial infarction). Specific units of measurement (formatted according to UCUM [45]) may also be added as properties so that units can be assigned to a numeric restriction and thus transmitted to the FHIR server (an example query with the unit "mmHG" is shown above). When FHIR queries are executed by the FHIR server, all matching resources (e.g., observations) are returned regardless of when they were acquired. Therefore, it may be useful to specify periods of time where relevant observations took place (e.g., last year). Additionally, to handle multiple occurrences of matching observations,

experts may add aggregation functions (minimum, average, first, etc.). Lastly, phenotype classes can be annotated as inclusion or exclusion criteria, which will affect how PhenoMan processes the FHIR query results for these classes.

By using the phenotype classes in a PheSO and the properties described above, PhenoMan can generate FHIR Search queries that are subsequently sent to an FHIR server via GET requests. The query results are then used to create reports on patients with matching conditions. The process can be broken down into the following steps. First, PhenoMan prepares an initial patient set, which is then further refined. This patient set is derived from master data constraints such as FHIR administrative gender [46], date of birth, or age ranges. Age ranges are converted into dates. An initial FHIR Search query may look like "Patient?birthdate=gt1954-04-24&birthdate=le1980-04-24&gender=male", and the result is a set of FHIR Patient resources. After the initial set of patients is obtained, PhenoMan generates corresponding queries for inclusion and exclusion criteria (see the examples in Table 1). The queries are sent to the FHIR server, and in the case of inclusion criteria, the results are combined with the initial patient set (set–theoretical intersection), whereas exclusion criteria results are removed from the initial patient set (set–theoretical difference).

Table 1. Example FHIR Search queries for the inclusion and exclusion criteria listed in Section 2.4. The query execution timestamp is assumed to be 13 May 2022 09:00:00. Here, "[base]" is a placeholder for an FHIR server uniform resource location (URL).

Criteria from Section 2.4	Resulting FHIR Search Queries
Inclusions:	
(ic1) male subjects	GET [base]/Patient?gender=male &birthdate=gt1956-05-13&birthdate=le1982-05-13
(ic2) age between 40 and 65 years (i.e., age in years $\geq$ 40 and <66)	
(ic3a) BMI $\geq$ 30 kg/m^2	GET [base]/Observation?code=http://loinc.org|3137-7
(requires height and weight)	GET [base]/Observation?code=http://loinc.org|3141-9
(ic3b) at least one observation of systolic blood pressure $\geq$130 mmHg in the last month	GET [base]/Observation ?component-code-value-quantity=http://loinc.org|8480-6 $ge130|http://unitsofmeasure.org|mm[Hg] &date=ge2022-04-13T09:00:00
Exclusions:	GET [base]/Condition
(ec1) myocardial infarction	?code=http://snomed.info/sct|22298006,
(ec2) stroke	http://snomed.info/sct|230690007

The results of (ic3a) and (ic3b) are combined in (ic3) (set–theoretical union).

A more detailed description of FHIR query building and the result set handling process is presented in [23,44].

3.1.2. Use of Study Data Query Language in the Leipzig Health Atlas

It is important to not only model phenotypes in PheSOs but also represent them in a user-friendly way (e.g., on websites) and make them browsable by experts from different disciplines (e.g., epidemiologists or physicians). This increases the visibility and reusability of these phenotypes (according to FAIR data principles). The phenotype class representations can be used to run simple phenotypic queries directly on research data management systems (RDMSs) with anonymised patient data. This would allow experts to retrieve case numbers to obtain an overview of the available research data without having to build complex queries in the RDMS native language.

A first prototypic approach [47] of the above-described connection between phenotype class representations and an RDMS was implemented in the Leipzig Health Atlas project [48]. One outcome of the project is a repository (also called the Leipzig Health Atlas (LHA)) for storing heterogeneous scientific projects, associated publications, results (e.g., (bio)medical and clinical data sets, models, and tools), and their metadata [49]. The LHA follows the FAIR data principles and was built on the SEEK management platform [50], which provides the basis for a repository with archiving, presentation, and secure sharing

capabilities. We have extended SEEK with methods to represent the phenotype classes of uploaded PheSOs. Another component of the LHA is the LHA Data Portal for managing study data and metadata and for querying data by user-specified filter criteria. The data portal is based on the LIFE Data Portal [51] (closed source, developed in the LIFE Study, Leipzig, Germany), and it uses the vendor-neutral and platform-independent CDISC ODM format for exchanging and archiving clinical and translational study data [52] to enable efficient and flexible storage and provision of study data in a standardised manner. Each query created with the data portal web client is internally translated into the custom domain-specific language Study Data Query Language (SDQL) [53]. SDQL reuses the conceptual abstract CDISC ODM entities to utilise known and well-defined terminologies.

We have developed an extension to the PhenoMan Core API that imports phenotype classes with pre-generated SDQL queries into the LHA. When a user browses the phenotype classes, they can execute the queries by clicking on a link to the data portal so that the corresponding query is executed, and the user is given the number of cases. To generate SDQL queries in advance, the PhenoMan extension transforms phenotype classes as follows. All single phenotype classes must be mapped to CDISC ODM items by annotating them with respective item OIDs (CDISC ODM-related identifications for study items such as measurements, laboratory values, and questions). We used the annotation property "alias" to realise the mapping. The simplest queries result from unrestricted single phenotype classes (e.g., sex or waist circumference) and are intended to query the number of study participants having a value (database entry) of the corresponding study items. Value range queries (e.g., waist circumference is greater than 120 cm) are generated from restricted single phenotype classes (e.g., "high waist circumference") and their value restrictions and return the number of study participants with study item values matching the range restriction. Lastly, Boolean expression queries are generated based on general class axioms describing the restricted combined phenotype classes (e.g., "has_part some male and has_part some high_waist_circimference" for substantially increased risk of metabolic complications). The resulting presentation of the described phenotype classes is available at [54] (login is required to execute queries on the LHA data portal).

3.1.3. Use of SQL in the LIFE Research Study

LIFE is an epidemiological and multi-cohort study conducted at the Leipzig Research Centre for Civilization Diseases (Leipzig University) [55,56]. The aim of LIFE is to determine the prevalence and causes of common civilisation diseases such as obesity, depression, and dementia by investigating thousands of Leipzig inhabitants of different ages. All participants are examined in a (multi-day) program using a wide range of assessments [56,57]. The assessments include interviews, self-completed questionnaires, instrumental examinations such as anthropometry, ECG, or MRI, and laboratory analyses of the collected samples. The data for each assessment, depending on the participant's examination program, are collected using special input systems and prepared entry forms. All collected data are integrated and harmonised in a central research database. This database consists of data tables related to assessments (i.e., investigations). The complexity of the tables (i.e., number of columns) varies widely. All instance data in the research database are described by metadata, which are composed of the table and column names, the corresponding data types, and the original question or description of the measurement and a possible code list. The metadata are stored in a dedicated metadata repository (MDR) and are linked to the instance data.

The collected data are analysed in an increasing number of analysis projects. Each project is specified by a project agreement that describes the analysis goal, plan, and required data. Current project agreements typically require data from up to 50 assessments. The queries to retrieve the data can be very complex. They typically combine data from multiple research database tables, multiple selection expressions, and a variety of projected columns. Manually specifying such database queries for each analysis project would be very error-prone and time-consuming. To solve this problem, we developed

an ontological framework for describing the metadata and an ontology-based tool, the LIFE Query Generator, which generates the required database queries from the ontological specification [57].

The ontological framework [57] consists of three interrelated layers. The integrated data layer includes all research data (instance data) from the research database. The metadata layer describes the instance data using the metadata specified in the MDR. Finally, the ontology layer is represented by the LIFE Investigation Ontology (LIO) and its mapping to the collected metadata in the MDR. LIO is the core component of the framework and a predecessor version of the COP. It is used as a general model to formulate phenotypic knowledge and generate SQL queries. LIO semantically classifies biomedical investigations in LIFE. Fundamentally, we differentiate between items (e.g., questions of a questionnaire or single laboratory parameters) and item sets (e.g., complete assessments). Item sets correspond to the data tables (or views) of the research database, whereas items are represented by their columns. The mappings between the collected metadata in the MDR (e.g., table and column names) and LIO categories are inherently generated.

The LIFE Query Generator [57] transforms the ontological specifications to corresponding SQL statements. Using the LIFE Query Generator, the applicant can search for desired assessments and select the complete assessments or specific items of an assessment. These selections are used to create the projection (i.e., items for which data should be retrieved). Furthermore, inclusion and exclusion criteria can be specified at the item level. Per default, the query generator creates one query for each selected assessment or item set. In addition, experienced users can create new item sets containing items from different assessments. Such item sets result in join queries, using patient identifiers and examination time points as join criteria. The SQL queries are then generated and executed in the research database. The corresponding research data are provided via a web-based reporting application.

In future work, we will adapt the SQL Query Generator to the current version of the COP.

4. Discussion

In this work, we present a novel approach for modelling the knowledge required for creating phenotypic queries. We prototypically implemented this approach for several query languages, such as FHIR and SQL. We recommend the use of Phenotype Specification Ontologies as a standardised method of modelling and managing phenotype definitions. As described in the methods section, these ontologies can serve as a first step towards FAIRification of phenotypes and ontologies (i.e., bringing them in line with the FAIR Data Principles) and thus improve reusability by domain experts. Currently, not all aspects of FAIR are covered because, on the one hand, ontological modelling of phenotypes is a very flexible approach and may result in incomplete ontologies that lack essential annotation properties such as references to controlled vocabularies, definitions, or even synonyms. On the other hand, the resulting ontologies and their metadata are not necessarily accessible to the community, as they still need to be made available via a web service. Of course, there are many services that are suitable for this aspect (e.g., the SEEK research data management platform [50]), but it may become cumbersome to build ontologies and publish them in different tools. Therefore, in our future work, we want to focus on building a comprehensive platform with integrated tools to create, search, publish, and reuse phenotype ontologies. In this platform, we will implement extended reasoning capabilities as well as query types that are currently not supported by PhenoMan.

The specification of search queries is usually performed by medical scientists (domain experts) who do not routinely use the respective query language of the source system. The effort required by scientists to translate clinical trial eligibility criteria into search queries, identifying the cohort of interest, can be substantial [3], and remote data centres with highly individual query languages further increase the time required to adopt queries for all centres. To reduce the number of query languages that scientists need to master, we introduced the Phenotype Specification Ontologies, which use a common model and can be

automatically mapped to the query languages of all data centre source systems. Mappings from a PheSO to the actual query languages are required and must be implemented. The amount of work required for the implementation depends highly on the source system, but it only needs to be performed once, and the resulting mapping is hidden from scientists. Admittedly, query languages can be very heterogeneous [10], and some may have less complex syntaxes than PheSOs and are easier to learn, but learning only modelling phenotypic knowledge (supported by suitable tools) still outweighs the higher complexity.

Van Spall et al. found that the exclusion criteria of randomised clinical trials are not always clearly reported in scientific publications [11]. They also noted that multicentric trials tend to have extensive exclusions and thus may impair the generalisability. These aspects show that there is a need for standardised representations of eligibility criteria. Our ontological approach could serve as one way to express the eligibility criteria of clinical trials and, in addition, allow execution in multiple data centres. Not all criteria can be expressed using logic-based languages [10], but Description Logic (including OWL2, which we are using for the PheSO) especially offers the advantage of class-based reasoning and annotation with rich metadata.

When building and executing phenotypic queries, it is important to consider that the result sets highly depend on the underlying data quality, such as in electronic health records. Unfortunately, there are many challenges when dealing with EHRs. Data quality must be considered when building queries because data may be missing, inaccurately documented, or biased [58–60]. For example, some centres participating in a clinical trial might rely on LOINC codes to identify observations, while others use SNOMED CT. Therefore, codes from multiple taxonomies must be added to the properties of a phenotype class to make sure that resulting queries are executable in all centres. The values of numeric phenotypes used in formulas of unrestricted derived phenotype classes may be missing, resulting in an incalculable phenotype. In this case, additional expressions (e.g., conditional if-else expressions) can be inserted into the formula to provide, for example, a default value or an alternative phenotype to be used as a substitute. Ultimately, only some aspects of poor data quality can be addressed at the level of phenotypic abstraction.

Nelson et al. proposed a low-burden, multicentric model for cohort assessments based on computable phenotype algorithms [61]. Their premise is that all participating centres are autonomous, diverse and operate individual solutions for count estimations (e.g., i2b2 [62], TriNetX [63], and direct database access). Phenotype algorithms are developed and discussed by informaticians and data scientists and shared between the centres so that they can run queries in their respective solutions, and the resulting count estimations are returned. We are confident that our solution can also be used in the setting described. For phenotypic queries resulting from PheSOs to be executable in all centres, appropriate transformations need to be implemented. After this initial overhead, phenotypic query execution is narrowed down to the exchange of PheSOs between centres. In addition, PheSOs can be developed by scientists without a background in informatics.

An ontological approach for building computable eligibility criteria commonly used in Hepatitis C virus clinical trials was proposed by Zhang et al. [3]. Their approach was based on the OBDA framework Ontop, which enables querying relational databases as virtual RDF graphs with SPARQL queries. They constructed an ontology for computable eligibility criteria in the Hepatitis C virus (OCED-HCV) based on the Basic Formal Ontology as the upper-level ontology and additional ontologies such as the Human Disease Ontology and the Ontology for Biomedical Investigations. Via semantic mapping axioms, entities in the OCED-HCV were linked to data constructs in the trial data management system. Researchers can use a web frontend to construct eligibility criteria and visualise cohort results. The frontend transmits the criteria to a backend, where they are translated into SPARQL queries and executed on the trial data. We believe that our approach of modelling phenotype classes and automatically constructing phenotypic queries is more generic than the one proposed by Zhang et al., because we did not prepare a disease-specific ontology but encouraged researchers in building phenotype classes and enriching them

with annotations on their own. The resulting standardised ontological phenotype models (PheSOs) can be shared, are usable independent of the trial data management system (with the drawback that mappings need to be implemented), and are not restricted to a limited set of diseases (in this case, entities in OCED-HCV).

The main limitation of our proposed ontology-based phenotypic model is that we cannot take advantage of the full expressive power of specific query languages (such as SQL or SPARQL). However, our approach is intended for domain experts, so we focused on the most relevant query types (phenotypic queries) that they use in their work. By providing user-friendly tools, domain experts can be efficiently supported in specifying phenotypic knowledge and building respective queries. Another limitation is the need to implement adapters for the corresponding data sources. We will explore solutions to simplifying adapter implementation.

5. Conclusions

We present a novel and generic approach to modelling phenotypic knowledge and automatically creating system-specific queries for data sources holding study and patient data. The queries are generated from standardised Phenotype Specification Ontologies, where phenotypes are modelled according to the general model provided by the Core Ontology of Phenotypes. The implementation of this approach enables domain experts to model phenotype classes and use them as eligibility criteria for a data source without having to know the query language of the data source.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/app12105214/s1. File S1: Example PheSO in OWL format, containing phenotypes for an example study with hypertension or obesity as inclusion criteria.

Author Contributions: Conceptualisation, A.U., C.B. and T.K.; methodology, A.U., C.B., H.H. and T.K.; validation, C.B., A.U., F.M. and R.S.; writing—original draft preparation, C.B., A.U. and T.K.; writing—review and editing, F.M., H.H. and R.S.; visualisation, A.U.; supervision, A.U.; project administration, A.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the German Federal Ministry of Education and Research as part of the projects SMITH TOP (grant number: 01ZZ2018) and SMITH (grand number: 01ZZ1803A).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in supplementary material S1 and are publicly available here: https://health-atlas.de/lha/8ATHQGDE4F-4 (accessed on 3 May 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nathan, R.A. How Important Is Patient Recruitment in Performing Clinical Trials? *J. Asthma* **1999**, *36*, 213–216. [CrossRef] [PubMed]
2. Sullivan, J. Subject Recruitment and Retention: Barriers to Success. *Applied Clinical Trials*, 1 April 2004.
3. Zhang, H.; He, Z.; He, X.; Guo, Y.; Nelson, D.R.; 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*, 1601–1610. [PubMed]
4. Obeid, J.S.; Beskow, L.M.; Rape, M.; Gouripeddi, R.; Black, R.A.; Cimino, J.J.; Embi, P.J.; Weng, C.; Marnocha, R.; Buse, J.B.; et al. A Survey of Practices for the Use of Electronic Health Records to Support Research Recruitment. *J. Clin. Trans. Sci.* **2017**, *1*, 246–252. [CrossRef] [PubMed]
5. Thadani, S.R.; Weng, C.; Bigger, J.T.; Ennever, J.F.; Wajngurt, D. Electronic Screening Improves Efficiency in Clinical Trial Recruitment. *J. Am. Med. Inform. Assoc.* **2009**, *16*, 869–873. [CrossRef]
6. Scheuermann, R.H.; Ceusters, W.; Smith, B. Toward an Ontological Treatment of Disease and Diagnosis. *Summit Transl. Bioinform.* **2009**, *2009*, 116–120.

7. Chapman, M.; Mumtaz, S.; Rasmussen, L.V.; Karwath, A.; Gkoutos, G.V.; Gao, C.; Thayer, D.; Pacheco, J.A.; Parkinson, H.; Richesson, R.L.; et al. Desiderata for the Development of Next-Generation Electronic Health Record Phenotype Libraries. *GigaScience* **2021**, *10*, giab059. [CrossRef]
8. Richesson, R.; Smerek, M.; Cameron, C.B. A Framework to Support the Sharing and Re-Use of Computable Phenotype Definitions Across Health Care Delivery and Clinical Research Applications. *eGEMs* **2016**, *4*, 2. [CrossRef]
9. Spratt, S.E.; Pereira, K.; Granger, B.B.; Batch, B.C.; Phelan, M.; Pencina, M.; Miranda, M.L.; Boulware, E.; Lucas, J.E.; Nelson, C.L.; et al. Assessing Electronic Health Record Phenotypes against Gold-Standard Diagnostic Criteria for Diabetes Mellitus. *J. Am. Med. Inform. Assoc.* **2017**, *24*, e121–e128. [CrossRef]
10. Weng, C.; Tu, S.W.; Sim, I.; Richesson, R. Formal Representation of Eligibility Criteria: A Literature Review. *J. Biomed. Inform.* **2010**, *43*, 451–467. [CrossRef]
11. Van Spall, H.G.C.; Toren, A.; Kiss, A.; Fowler, R.A. Eligibility Criteria of Randomized Controlled Trials Published in High-Impact General Medical Journals: A Systematic Sampling Review. *JAMA* **2007**, *297*, 1233. [CrossRef] [PubMed]
12. SPARQL Query Language for RDF. Available online: https://www.w3.org/TR/rdf-sparql-query (accessed on 25 March 2022).
13. Arden Syntax | HL7. Available online: https://www.hl7.org/special/Committees/arden/index.cfm (accessed on 15 February 2022).
14. Heflin, J.; Hendler, J. A Portrait of the Semantic Web in Action. *IEEE Intell. Syst.* **2001**, *16*, 54–59. [CrossRef]
15. Poggi, A.; Lembo, D.; Calvanese, D.; De Giacomo, G.; Lenzerini, M.; Rosati, R. Linking Data to Ontologies. In *Journal on Data Semantics X*; Spaccapietra, S., Ed.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2008; Volume 4900, pp. 133–173, ISBN 978-3-540-77687-1.
16. Heymans, S.; Ma, L.; Anicic, D.; Ma, Z.; Steinmetz, N.; Pan, Y.; Mei, J.; Fokoue, A.; Kalyanpur, A.; Kershenbaum, A.; et al. Ontology Reasoning with Large Data Repositories. In *Ontology Management*; Hepp, M., Leenheer, P., Moor, A., Sure, Y., Eds.; Computing for Human Experience; Springer: Boston, MA, USA, 2008; Volume 7, pp. 89–128. ISBN 978-0-387-69899-1.
17. Mo, H.; Thompson, W.K.; Rasmussen, L.V.; Pacheco, J.A.; Jiang, G.; Kiefer, R.; Zhu, Q.; Xu, J.; Montague, E.; Carrell, D.S.; et al. Desiderata for Computable Representations of Electronic Health Records-Driven Phenotype Algorithms. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 1220–1230. [CrossRef] [PubMed]
18. Mahner, M.; Kary, M. What Exactly Are Genomes, Genotypes and Phenotypes? And What About Phenomes? *J. Theor. Biol.* **1997**, *186*, 55–63. [CrossRef]
19. Hoehndorf, R.; Oellrich, A.; Rebholz-Schuhmann, D. Interoperability between Phenotype and Anatomy Ontologies. *Bioinformatics* **2010**, *26*, 3112–3118. [CrossRef]
20. Uciteli, A.; Groß, S.; Kireyev, S.; Herre, H. An Ontologically Founded Architecture for Information Systems in Clinical and Epidemiological Research. *J. Biomed. Sem.* **2011**, *2*, S1. [CrossRef]
21. Bucur, A.; van Leeuwen, J.; Chen, N.-Z.; Claerhout, B.; de Schepper, K.; Perez-Rey, D.; Paraiso-Medina, S.; Alonso-Calvo, R.; Mehta, K.; Krykwinski, C. Cohort Selection and Management Application Leveraging Standards-Based Semantic Interoperability and a Groovy DSL. In *AMIA Summits on Translational Science Proceedings*; American Medical Informatics Association: Bethesda, MD, USA, 2016; Volume 2016, pp. 25–32.
22. Robinson, P.N. Deep Phenotyping for Precision Medicine. *Hum. Mutat.* **2012**, *33*, 777–780. [CrossRef]
23. Uciteli, A.; Beger, C.; Kirsten, T.; Meineke, F.A.; Herre, H. Ontological Representation, Classification and Data-Driven Computing of Phenotypes. *J. Biomed. Semant.* **2020**, *11*, 15. [CrossRef]
24. Waist Circumference and Waist-Hip Ratio: Report of a WHO Expert Consultation. Available online: https://www.who.int/publications/i/item/9789241501491 (accessed on 18 February 2022).
25. Musen, M.A. The Protégé Project: A Look Back and a Look Forward. *AI Matters* **2015**, *1*, 4–12. [CrossRef]
26. Winter, A.; Stäubert, S.; Ammon, D.; Aiche, S.; Beyan, O.; Bischoff, V.; Daumke, P.; Decker, S.; Funkat, G.; Gewehr, J.; et al. Smart Medical Information Technology for Healthcare (SMITH): Data Integration Based on Interoperability Standards. *Methods Inf. Med.* **2018**, *57*, e92–e105. [CrossRef]
27. Type 2 Diabetes Mellitus | PheKB. Available online: https://www.phekb.org/phenotype/type-2-diabetes-mellitus (accessed on 4 January 2022).
28. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]
29. Mons, B.; Neylon, C.; Velterop, J.; Dumontier, M.; da Silva Santos, L.O.B.; Wilkinson, M.D. Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud. *Inf. Serv. Use* **2017**, *37*, 49–56. [CrossRef]
30. DCMI Metadata Terms. Available online: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ (accessed on 4 January 2022).
31. PROV-O: The PROV Ontology. Available online: https://www.w3.org/TR/2013/REC-prov-o-20130430/ (accessed on 4 January 2022).
32. OWL 2 Web Ontology Language (Second Edition). Available online: https://www.w3.org/TR/owl2-overview/ (accessed on 4 January 2022).

33. Wilkinson, M.D.; Verborgh, R.; da Silva Santos, L.O.B.; Clark, T.; Swertz, M.A.; Kelpin, F.D.L.; Gray, A.J.G.; Schultes, E.A.; van Mulligen, E.M.; Ciccarese, P.; et al. Interoperability and FAIRness through a Novel Combination of Web Technologies. *PeerJ Comput. Sci.* **2017**, *3*, e110. [CrossRef]
34. Bernstam, E.V.; Warner, J.L.; Krauss, J.C.; Ambinder, E.; Rubinstein, W.S.; Komatsoulis, G.; Miller, R.S.; Chen, J.L. Quantitating and Assessing Interoperability between Electronic Health Records. *J. Am. Med Inf. Assoc* **2022**, *29*, 753–760. [CrossRef]
35. Halamka, J.D.; Tripathi, M. The HITECH Era in Retrospect. *N. Engl. J. Med.* **2017**, *377*, 907–909. [CrossRef]
36. Gehring, S.; Eulenfeld, R. German Medical Informatics Initiative: Unlocking Data for Research and Health Care. *Methods Inf. Med.* **2018**, *57*, e46–e49. [CrossRef]
37. Semler, S.; Wissing, F.; Heyder, R. German Medical Informatics Initiative: A National Approach to Integrating Health Data from Patient Care and Medical Research. *Methods Inf. Med.* **2018**, *57*, e50–e56. [CrossRef]
38. Der Kerndatensatz Der Medizininformatik-Initiative. Available online: https://www.medizininformatik-initiative.de/de/der-kerndatensatz-der-medizininformatik-initiative (accessed on 4 January 2022).
39. HL7 FHIR v4.0.1. Available online: https://www.hl7.org/fhir/ (accessed on 4 January 2022).
40. FHIR Search v4.0.1. Available online: https://www.hl7.org/fhir/search.html (accessed on 4 January 2022).
41. HAPI FHIR—The Open Source FHIR API for Java. Available online: https://hapifhir.io/ (accessed on 4 January 2022).
42. Vonk. Available online: https://www.gefyra.de/p/vonk.html (accessed on 7 January 2022).
43. Blaze: A FHIR Server with Internal, Fast CQL Evaluation Engine. Available online: https://github.com/samply/blaze (accessed on 4 January 2022).
44. Uciteli, A.; Beger, C.; Wagner, J.; Kirsten, T.; Meineke, F.A.; Stäubert, S.; Löbe, M.; Herre, H. Ontological Modelling and FHIR Search Based Representation of Basic Eligibility Criteria. *GMS Med. Inform. Biometry Epidemiol.* **2021**, *17*, Doc05. [CrossRef]
45. UCUM: The Unified Code for Units of Measure. Available online: https://ucum.org/trac (accessed on 6 January 2022).
46. Codesystem-Administrative-Gender—FHIR v4.0.1. Available online: https://www.hl7.org/fhir/codesystem-administrative-gender.html (accessed on 6 January 2022).
47. Uciteli, A.; Beger, C.; Wagner, J.; Kiel, A.; Meineke, F.A.; Stäubert, S.; Löbe, M.; Hänsel, R.; Schuster, J.; Kirsten, T.; et al. Ontological Modelling and Execution of Phenotypic Queries in the Leipzig Health Atlas. In *Studies in Health Technology and Informatics*; IOS Press: Amsterdam, The Netherlands, 2021; Volume 278, pp. 66–74, ISBN 978-1-64368-176-4.
48. Meineke, F.A.; Löbe, M.; Stäubert, S. Introducing Technical Aspects of Research Data Management in the Leipzig Health Atlas. *Stud. Health Technol. Inf.* **2018**, *247*, 426–430.
49. The Leipzig Health Atlas. Available online: https://health-atlas.de/ (accessed on 7 January 2022).
50. Wolstencroft, K.; Owen, S.; Krebs, O.; Nguyen, Q.; Stanford, N.J.; Golebiewski, M.; Weidemann, A.; Bittkowski, M.; An, L.; Shockley, D.; et al. SEEK: A Systems Biology Data and Model Management Platform. *BMC Syst. Biol.* **2015**, *9*, 33. [CrossRef]
51. Kirsten, T.; Kiel, A.; Wagner, J.; Rühle, M.; Löffler, M. *Selecting, Packaging, and Granting Access for Sharing Study Data*; Gesellschaft für Informatik: Bonn, Germany, 2017; ISBN 978-3-88579-669-5.
52. CDISC ODM. Available online: https://www.cdisc.org/standards/data-exchange/odm (accessed on 7 January 2022).
53. Wagner, J. Softwaregestützte Bereitstellung von Epidemiologischen Forschungsdaten. Master's Thesis, Leipzig University of Applied Sciences, Leipzig, Germany, 2016.
54. Body Mass Index, Waist Circumference and Waist-to-Hip Ratio. Available online: https://www.health-atlas.de/phenotype_algorithms/BMI_Waist_Hip (accessed on 7 January 2022).
55. LIFE—Leipziger Forschungszentrum Für Zivilisationserkrankungen. Available online: https://www.uniklinikum-leipzig.de/einrichtungen/life (accessed on 5 January 2022).
56. Loeffler, M.; Engel, C.; Ahnert, P.; Alfermann, D.; Arelin, K.; Baber, R.; Beutner, F.; Binder, H.; Brähler, E.; Burkhardt, R.; et al. The LIFE-Adult-Study: Objectives and Design of a Population-Based Cohort Study with 10,000 Deeply Phenotyped Adults in Germany. *BMC Public Health* **2015**, *15*, 691. [CrossRef] [PubMed]
57. Uciteli, A.; Kirsten, T. Ontology-Based Retrieval of Scientific Data in LIFE. In *Datenbanksysteme für Business, Technologie und Web (BTW 2015)-Workshopband*; Gesellschaft für Informatik: Bonn, Germany, 2015.
58. Hripcsak, G.; Albers, D.J. Next-Generation Phenotyping of Electronic Health Records. *J. Am. Med. Inform. Assoc.* **2013**, *20*, 117–121. [CrossRef] [PubMed]
59. Weiskopf, N.G.; Weng, C. Methods and Dimensions of Electronic Health Record Data Quality Assessment: Enabling Reuse for Clinical Research. *J. Am. Med. Inform. Assoc.* **2013**, *20*, 144–151. [CrossRef] [PubMed]
60. Weiskopf, N.G.; Rusanov, A.; Weng, C. Sick Patients Have More Data: The Non-Random Completeness of Electronic Health Records. *AMIA Annu. Symp. Proc.* **2013**, *2013*, 1472–1477. [PubMed]
61. Nelson, S.J.; Drury, B.; Hood, D.; Harper, J.; Bernard, T.; Weng, C.; Kennedy, N.; LaSalle, B.; Gouripeddi, R.; Wilkins, C.H.; et al. EHR-Based Cohort Assessment for Multicenter RCTs: A Fast and Flexible Model for Identifying Potential Study Sites. *J. Am. Med. Inform. Assoc.* **2021**, *29*, 652–659. [CrossRef]
62. Murphy, S.; Wilcox, A. Mission and Sustainability of Informatics for Integrating Biology and the Bedside (I2b2). *EGEMS* **2014**, *2*, 1074. [CrossRef]
63. Topaloglu, U.; Palchuk, M.B. Using a Federated Network of Real-World Data to Optimize Clinical Trials Operations. *JCO Clin. Cancer Inf.* **2018**, *2*, 1–10. [CrossRef]