Model-driven execution of phenotype algorithms – introduction of the Terminology- and Ontology-based Phenotyping Framework

Modell-getriebene Ausführung von Phänotyp-Algorithmen – Einführung des Terminology- and Ontology-based Phenotyping Framework

Christoph Beger[1,2]
Franz Matthies[1]
Ralph Schäfermeier[1]
Alexandr Uciteli[1]

1 Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University, Leipzig, Germany

2 Growth Network CrescNet, Leipzig University, Leipzig, Germany

Abstract

Automated detection of patients who meet certain clinical criteria or who experience adverse events plays a key role in improving patient care. Rule-based systems are one approach to enable standardised modelling of algorithms to detect patients with certain observable characteristics (phenotypes). Another important aspect is the retrieval of data from electronic health records in hospital information systems and clinical research database systems. In this work, we propose a new rule-based system, suitable for domain experts without IT background that allows automatic retrieval of patient data on execution by translating search queries into source-specific query languages.

As a basis for rule-based structured phenotype algorithms, we use a model derived from the Core Ontology of Phenotypes. The model allows phenotype algorithms to be defined with properties such as code system relationships to improve interoperability. With a relatively simple mapping, IT specialists can provide a specification to connect these models to a local hospital information system.

We developed an interactive web application, the Terminology- and Ontology-based Phenotyping Framework, to support domain experts in modelling and executing phenotype algorithms. Depending on the data-holding source system, an appropriate adapter is used to translate and execute queries in the source system-specific language. The framework was tested against a SQL database and a FHIR server, both initialised with randomly generated data. Generated queries yielded identical result sets.

The proposed computer-assisted approach can improve disease detection and clinical trial recruitment and enables a clear division of tasks between domain experts and IT specialists.

Keywords: phenotype, algorithms, eligibility determination, health information interoperability, terminology as topic

Zusammenfassung

Die automatisierte Erkennung von Patienten, die bestimmte klinische Kriterien erfüllen oder unerwünschte Ereignisse aufweisen, spielt eine wichtige Rolle bei der Verbesserung der Patientenversorgung. Regelbasierte Systeme sind eine Möglichkeit, die standardisierte Modellierung von Algorithmen zur Erkennung von Patienten mit bestimmten beobachtbaren Merkmalen (Phänotypen) zu ermöglichen. Ein weiterer wichtiger Aspekt ist der Abruf von Daten aus elektronischen Patientenakten in Krankenhausinformationssystemen und klinischen Forschungsdatenbanksystemen. In dieser Arbeit wird ein neues regelbasiertes System beschrieben, das für Fachleute ohne IT-Hintergrund geeignet ist und die automatische Abfrage von Patientendaten bei der Ausführung ermöglicht, indem Suchanfragen in quellenspezifische Abfragesprachen übersetzt werden.

Als Grundlage für regelbasierte strukturierte Phänotyp-Algorithmen wird ein Modell verwendet, das von der Core Ontology of Phenotypes abgeleitet ist. Das Modell ermöglicht die Definition von Phänotyp-Algorithmen mit Eigenschaften wie Code-System-Beziehungen zur Verbesserung der Interoperabilität. Mit einem relativ einfachen Mapping können IT-Spezialisten eine Spezifikation erstellen, um diese Modelle mit einem lokalen Krankenhausinformationssystem zu verbinden.

Wir haben eine interaktive Webanwendung, das Terminology- and Ontology-based Phenotyping Framework, entwickelt, um Fachleute bei der Modellierung und Ausführung von Phänotyp-Algorithmen zu unterstützen. Je nach dem datenhaltenden Quellsystem wird ein geeigneter Adapter verwendet, der die Abfragen in die quellsystemspezifische Sprache übersetzt und ausführt. Das Framework wurde mit einer SQL-Datenbank und einem FHIR-Server getestet, die beide mit zufällig generierten Daten initialisiert wurden. Die generierten Anfragen ergaben identische Ergebnismengen.

Der vorgeschlagene computergestützte Ansatz kann die Erkennung von Krankheiten und die Rekrutierung für klinische Studien verbessern und ermöglicht eine klare Aufgabenteilung zwischen Domänenexperten und IT-Spezialisten.

Introduction

In the Medical Informatics Initiative (MII) innovative IT solutions are developed to support medical research and enhance patient care. An important aspect of this is the automatic detection of diseases, risks for the same and adverse events of medications. For the implementation of such detection, patient data must be extracted from electronic health records (EHR) or research databases and evaluated. This may involve machine-interpretable phenotype algorithms, which use structured filter criteria and rules to detect individuals with specific properties (phenotypes) and derive further properties. With this, computer-assisted algorithms can, among other things, improve clinical trial recruitment [1]. The National Portal for Medical Research Data (FDPG) was recently developed in the MII [2]. It provides scientists with federated access to 31 German university hospitals, enabling them to perform feasibility queries and submit data use proposals. The query interface allows relatively simple matching criteria to be defined.

In this work, we want to focus on a more general approach of defining shareable models that can be used to perform complex queries and calculations in order to find matching individuals and infer additional characteristics. For this, we distinguish between three core notions: *phenotype*, *phenotype model* and *phenotype algorithm*.

We use the definition of '*phenotype*' as a '(combination of) bodily feature(s) of an organism determined by the interaction of its genetic make-up and environment' by Scheuermann et al. [3]. And in a broader sense, we also consider all observable human characteristics to be phenotypes.

A *phenotype model* is a formal representation of phenotypic knowledge. We hereby distinguish between general and specific phenotype models [4]. The general model provides a basic classification of phenotypes, the required attributes of phenotype classes (e.g., labels, data types, codes, units of measurement, derivation rules/formulas etc.), relation types and rules (axioms), which must be used/followed by developing specific models. A specific phenotype model represents phenotypic knowledge by defining specific phenotype classes (such as 'weight', 'height', 'BMI', 'obesity', etc.) according to the general model. This means, all required attributes of the classes (e.g., including the formula for calculating BMI from weight and height, as well as the ranges to classify the BMI value in categories from underweight to obesity) and relations between the classes (e.g., weight and height values are required for calculating BMI) are specified.

A *phenotype algorithm* is generally understood as an algorithm designed to identify or classify phenotypic traits [5], [6]. Specifically, we consider a phenotype algorithm to be a sequence of instructions to

1. query data representing phenotypes,
2. derive additional properties (derived/composite phenotypes) from other phenotypes,
3. classify phenotypes into phenotype classes, and
4. provide results in a machine- and human-readable format.

A *generic phenotype algorithm* is model-driven (i.e., it can be configured by a phenotype model). This means, the algorithm always performs the same (generic) steps (see above) and uses the specific phenotype definitions specified in a phenotype model (e.g., identifying obese patients based on weight and height or identifying patients with acute kidney injury based on creatinine).

The effort required to design algorithms can be very high depending on the complexity [7] and typically requires a close collaboration between domain and IT experts. To support domain experts in creating phenotype models and executing phenotype algorithms, we developed a framework, the Terminology- and Ontology-based Phenotyping (TOP) Framework, which will be presented in this work. The goal was for the framework to allow domain experts to define phenotypes with respective filters and roles on their own, and to be embeddable by IT experts into the hospital's information system, thus allowing the execution of algorithms on patient data from routine patient care and medical research.

Methods

We use the Core Ontology of Phenotypes (COP) [8] as the generic phenotype model and as a basis for developing specific models. COP specifies that all phenotypes of organisms are instances of one or more phenotype classes. These classes are distinguished into single and composite phenotype classes. Single phenotypes (e.g., height or weight of a person) are atomic and can be components/parts of composite phenotypes (e.g., BMI). The two described class types can further be subdivided into unrestricted and restricted phenotype classes. A restricted class (e.g., 'height above 150 cm') is always a subclass of an unrestricted class ('height') and possess a value restriction so that only a subset of the corresponding unrestricted class instances belongs to the restricted class. For example, the phenotype 'height' with value 180 cm of the individual John Doe is an instance of the unrestricted single phenotype class 'height' and, at the same time, of the restricted single phenotype class 'height above 150 cm'.

Single phenotype classes must be annotatable with terms of standard terminologies like Logical Observation Identifiers Names and Codes (LOINC) or Systematized Nomenclature of Medicine and Clinical Terms (SNOMED CT). These annotations are used to assign the phenotype classes with corresponding data from the source systems. For instance, the phenotype 'height' is annotated with code 'LOINC:3137-7'. With this annotation in place, a mapping can be provided to assign all height values from electronic health records to the phenotype 'height', which makes them available for classifications and calculations.

As composite phenotypes result from calculations and compositions of single phenotypes, constants and other composite phenotypes, we developed a generic specification to express these cases. According to the specification, all composite phenotype classes have exactly one expression representing either a phenotype, a constant, or exactly one function with a set of arguments. Expressions can be nested because arguments are also expressions. Functions convert argument sets into a single expression (usually a value), though they are not limited to mathematical functions and the available set of functions is extensible. For instance, the expression of the phenotype 'Body Mass Index' can be expressed as follows: 'quotient(weight, power(height, 2))'.

The TOP Framework implements a generic phenotype algorithm, which can be executed by selecting a desired phenotype model and defining (single or composite) phenotype classes as in- or exclusion criteria. Composite phenotypes can be traced back to singles from which they are derived. This results in a set of single phenotypes, which all have terminology annotations. For each of them, queries can be inferred that are translated to the respective query language of the source system and executed on the system. We assume that relevant data elements are available in structured form and are retrievable from data-holding systems (source system) with respective query languages. In the MII, Fast Healthcare Interoperability Resources (FHIR, https://www.hl7.org/fhir), by the standardisation organisation Health Level 7, is used to make patient data from electronic health records accessible. And also, many research data management systems are using databases based on the Structured Query Language (SQL). We therefore focused on SQL database management systems and FHIR servers with FHIR Search (https://www.hl7.org/fhir/search.html) support. The inference is done by generic or source system-specific adapters. For SQL and FHIR, we provide generic Java-based adapters based on [4], which are configurable by a mapping file. The query results are subsequently used to evaluate expressions of composite phenotypes.

The whole process of modelling phenotypes and executing phenotype algorithms is depicted in Figure 1 and can be carried out by domain experts. Only adapters and an optional source system mapping must be provided by IT specialists. The mapping associates the terminology codes of the phenotype classes in the phenotype model with the corresponding data elements contained in the source system, e.g. FHIR resource types like 'Observation' and 'Condition' or SQL database tables and columns. In addition, value ranges contained in the phenotype algorithm can be modified. This is especially helpful to customise an algorithm to facility-specific norm values like laboratory thresholds. If the terminology codes and also value ranges in the phenotype model and in the source system are identical, the mapping can be omitted.

Results

Based on the described methodological blueprint, we have developed an interactive web application, the TOP Framework (see Figure 2). The framework is used by domain experts to develop phenotype models and execute phenotype algorithms. It consists of a JavaScript-based frontend, a Java Spring Backend, and a vendor-independent database. The framework can optionally be connected to an Ontology Lookup Service (version 3) [9]. The main functions provided by the corresponding user interface elements or API are the following:

1. Developing phenotype models: Phenotype classes, attributes, relations, ranges, and expressions/formulas can be structurally defined.
2. Storing and exporting phenotype models: Phenotype models are stored in a relational database and can

be exported in different formats (e.g., Ontology Web Language, CSV or Decision Model and Notation).

3. Execution of phenotype algorithms: Phenotype algorithms can be initiated by selecting a phenotype model (in-/exclusion criteria) and a source system. Depending on the selected source system, a suitable adapter is used to translate and execute queries in source system-specific language.
4. Providing results: The number of patients/probands who fulfil all criteria of the phenotype model or also the corresponding data are provided in a machine and human readable format (e.g., CSV).

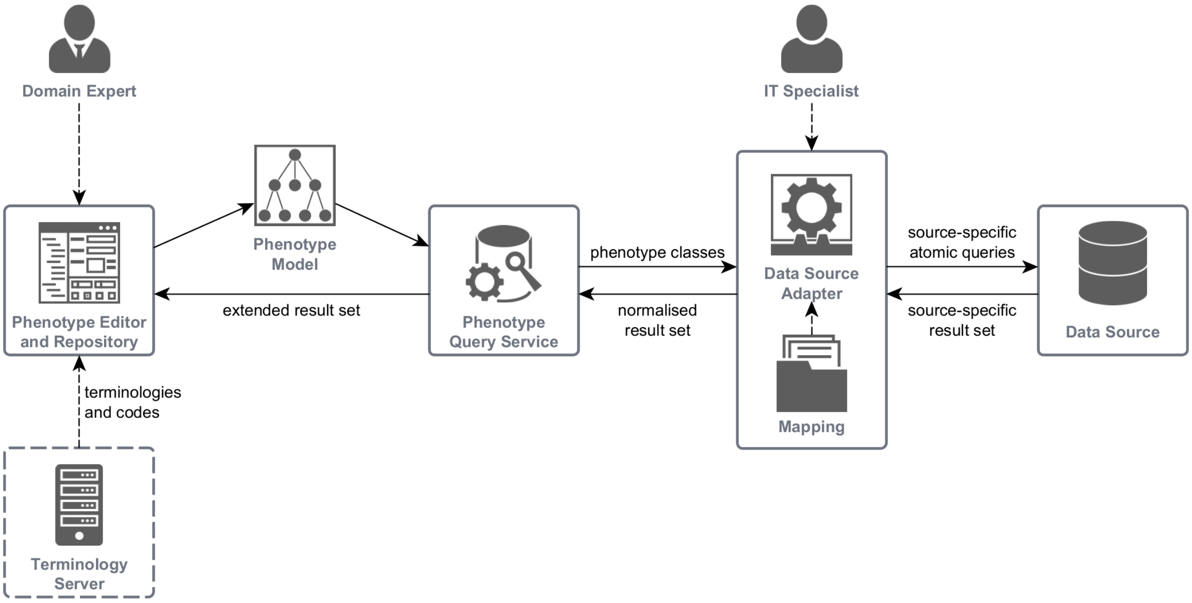
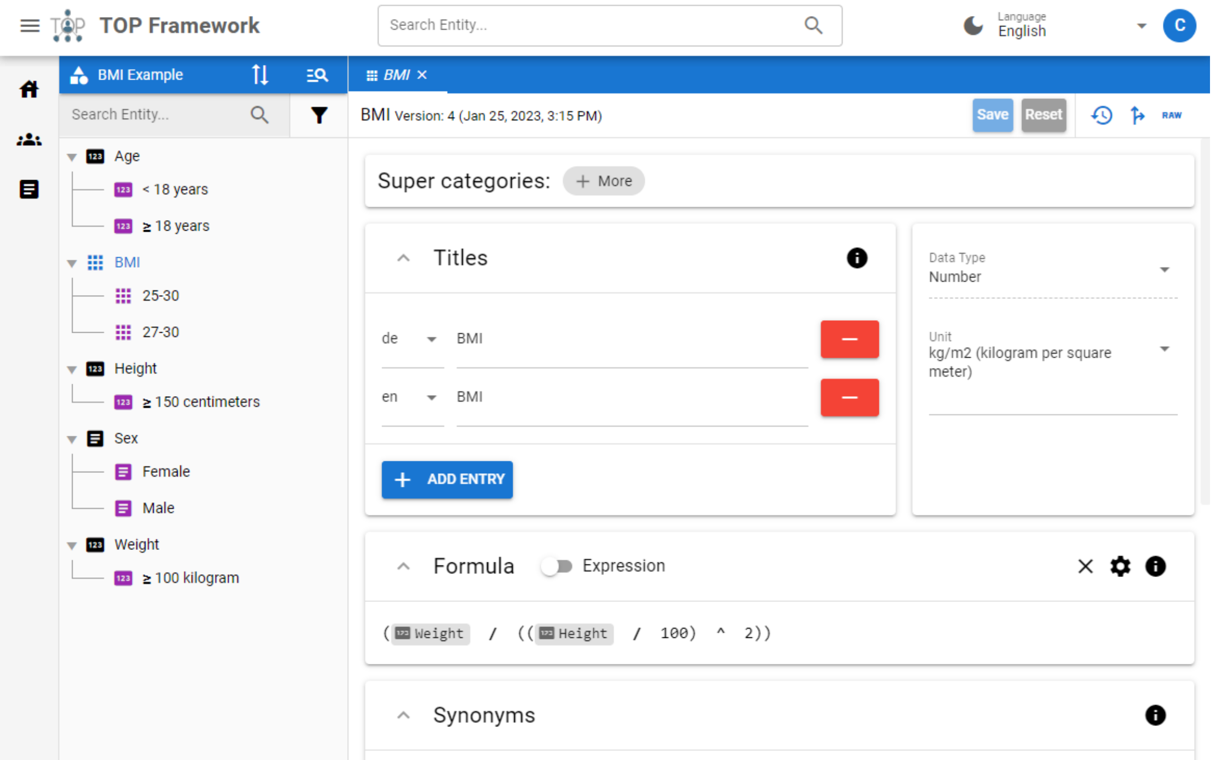


Figure 1: Overview of the TOP Framework. Domain experts use the graphical user interface to browse existing phenotype classes, create new ones and execute algorithms based on them. IT specialists provide data source adapters and mappings that enable generating source system-specific queries. External terminology servers can be incorporated.



- Age
  - < 18 years
  - ≥ 18 years
- BMI
  - 25-30
  - 27-30
- Height
  - ≥ 150 centimeters
- Sex
  - Female
  - Male
- Weight
  - ≥ 100 kilogram

BMI Version: 4 (Jan 25, 2023, 3:15 PM)

de BMI

en BMI

Data Type: Number

Unit: kg/m2 (kilogram per square meter)

```
( Weight  /  (( Height  /  100)  ^  2))
```

Figure 2: Screenshot of the TOP Framework web application. Shown is the repository ‘BMI Example’ that contains a simple phenotype model of the body mass index (BMI). On the left side, all phenotype definitions contained in the repository are displayed in a tree structure. And on the right side, a form to edit the definition of the phenotype ‘BMI’ is shown.

Furthermore, the framework allows search in existing phenotype models and supports, among other things, the creation of expressions for composite phenotypes and thus ensures syntactic correctness. Frontend and backend communicate over an API (the API's OpenAPI specification is available in [10]), which can also be used to integrate backend functionalities into other software applications.

The framework was tested with randomly generated test data consisting of 10,000 patients with about 50,000 visits. We provided the data both in an SQL database and via a FHIR server and created corresponding mapping configurations. Queries to both source systems yielded identical result sets.

Discussion

Typically, bioinformaticians, statisticians and comparable groups of people are tasked with implementing algorithms to detect individuals with specific traits. Since domain experts (i.e., medical staff) must provide the knowledge to implement such algorithms, very close collaboration between both groups is required [11]. Algorithms are often written in programming languages such as R, Python, Clinical Quality Language, which are more powerful and expressive than the TOP Framework but are not familiar to domain experts. As a result, there is no clear separation between modelling and implementation tasks. In addition, access to the source data must be implemented separately, e.g. using libraries such as FHIRcrackr [12] or FHIR-PYrate [13]. The TOP Framework can solve these issues by the described separation between phenotype models and phenotype algorithms. Consider phenotype models as different configurations applied to the generic phenotype algorithm described earlier. The framework provides a user interface suitable for domain experts to create such models and use them to compute or infer complex phenotypes for a given set of patient data, and to query patient data management systems for individuals with matching characteristics. Query results can be numbers of individuals or data sets containing selected patient characteristics for further statistical analysis. The approach of separating modelling and implementation tasks has been used previously by Uciteli et al. [14], [15], where ontological models built by domain experts were used to search text corpora and to identify perioperative risks.

The expressions used in the TOP Framework for modelling composite phenotypes are generic and thus suitable for mapping a large part, but not all, of the calculations and rules that occur in practice in algorithms. Some phenotypes can only be assessed using complex computations (e.g., Python and R scripts) or trained models (e.g., from machine learning). They are therefore not yet supported by the framework. In the future, we would like to integrate such complex calculations into the TOP Framework as external services.

The TOP Framework is a platform for easy definition of reusable phenotype models. All models can be accessed and referenced via URLs. However, the model is only partially standards compliant and introduces a proprietary format. To overcome this, we plan to extend the model with additional audit and metadata, and to enable automated transformation into community standard formats such as RDF with common annotations. This is in line with our commitment to the FAIR data principles [16].

The focus of our work was not to provide yet another query tool for patient and trial data, but to enable and encourage domain experts to build machine-readable and richly annotated models of phenotypes that can later be used to build queries. An initial overhead is required to build these models. Therefore, the framework does not support rapid query building as provided by tools such as FDPG [2], i2b2 [17], or tranSMART [18]. But there is no need for complex ETL jobs to pre-process patient or trial data. All queries are generated in the source-specific language and can be executed directly on the source system. In this work, we have focused on the query languages FHIR Search and SQL. However, the concept described here can also be used in conjunction with other query languages by implementing appropriate query adapters. Admittedly, the required mapping to source systems may be difficult for some database schemas or APIs – or even impossible in some cases if essential data is missing (e.g., terminology codes). However, the resulting framework is lightweight and has no requirements for the storage of patient data, and aspects such as data privacy can be handled by the source systems.

Conclusions

We have developed a methodology for modelling phenotypes and retrieving the corresponding data elements from patient data management systems and clinical research systems. The required tasks are clearly divided between domain experts and IT specialists. This computer-assisted approach can improve disease detection and clinical trial recruitment.

Notes

Terminology- and Ontology-based Phenotyping (TOP) Framework

Source code for frontend and backend is available at https://github.com/Onto-Med/top-frontend and https://github.com/Onto-Med/top-backend. A user manual can be found at https://onto-med.github.io/top-deployment.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the German Federal Ministry of Education and Research as part of the projects SMITH TOP (reference number: 01ZZ2018) and SMITH (reference number: 01ZZ1803A).

References

1. Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D. Electronic screening improves efficiency in clinical trial recruitment. J Am Med Inform Assoc. 2009;16(6):869-73. DOI: 10.1197/jamia.M3119
2. Prokosch HU, 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. Stud Health Technol Inform. 2023 May 18;302:307-11. DOI: 10.3233/SHTI230124
3. Scheuermann RH, Ceusters W, Smith B. Toward an ontological treatment of disease and diagnosis. Summit Transl Bioinform. 2009 Mar 1;2009:116-20.
4. Beger C, Matthies F, Schäfermeier R, Kirsten T, Herre H, Uciteli A. Towards an Ontology-Based Phenotypic Query Model. Appl Sci. 2022 May 21;12(10):5214. DOI: 10.3390/app12105214
5. Wei WQ, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Med. 2015;7(1):41. DOI: 10.1186/s13073-015-0166-y
6. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc. 2013 Dec;20(e2):e206-11. DOI: 10.1136/amiajnl-2013-002428
7. Zhang H, He Z, He X, Guo Y, Nelson DR, Modave F, Wu Y, Hogan W, Prosperi M, Bian J. Computable Eligibility Criteria through Ontology-driven Data Access: A Case Study of Hepatitis C Virus Trials. AMIA Annu Symp Proc. 2018 Dec 5;2018:1601-10.
8. Uciteli A, Beger C, Kirsten T, Meineke FA, Herre H. Ontological representation, classification and data-driven computing of phenotypes. J Biomed Semantics. 2020 Dec;11(1):15. DOI: 10.1186/s13326-020-00230-0
9. Jupp S, Burdett T, Leroy C, Parkinson HE. A new Ontology Lookup Service at EMBL-EBI. In: Malone J, Stevens R, Forsberg K, Splendiani A, editors. Proceedings of the 8th International Conference on Semantic Web Applications and Tools for Life Sciences (SWAT4LS 2015); 2015 Dec 7-10; Cambridge, UK. p. 118-9. URN: urn:nbn:de:0074-1546-1
10. Beger C, Matthies F. Onto-Med/top-api. Version v0.7.6. Zenodo; 2023 May 5. DOI: 10.5281/zenodo.7900530
11. Hruby GW, Boland MR, Cimino JJ, Gao J, Wilcox AB, Hirschberg J, Weng C. Characterization of the biomedical query mediation process. AMIA Jt Summits Transl Sci Proc. 2013 Mar 18;2013:89-93.
12. Palm J, Meineke FA, Przybilla J, Peschel T. “fhircrackr”: An R Package Unlocking Fast Healthcare Interoperability Resources for Statistical Analysis. Appl Clin Inform. 2023 Jan;14(1):54-64. DOI: 10.1055/s-0042-1760436
13. Hosch R, Baldini G, Parmar V, Borys K, Koitka S, Engelke M, Arzideh K, Ulrich M, Nensa F. FHIR-PYrate: a data science friendly Python package to query FHIR servers. BMC Health Serv Res. 2023 Jul;23(1):734. DOI: 10.1186/s12913-023-09498-1
14. Uciteli A, Kropf S, Weiland T, Meese S, Graef K, Rohrer S, Schurr MO, Bartussek W, Goller C, Blohm P, Seidel R, Bayer C, Kernenbach M, Pfeiffer K, Lauer W, Meyer JU, Witte M, Herre H. Ontology-based specification and generation of search queries for post-market surveillance. J Biomed Semantics. 2019 May;10(1):9. DOI: 10.1186/s13326-019-0203-7
15. Uciteli A, Neumann J, Tahar K, Saleh K, Stucke S, Faulbrück-Röhr S, Kaeding A, Specht M, Schmidt T, Neumuth T, Besting A, Stegemann D, Portheine F, Herre H. Ontology-based specification, identification and analysis of perioperative risks. J Biomed Semantics. 2017 Sep;8(1):36. DOI: 10.1186/s13326-017-0147-8
16. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. DOI: 10.1038/sdata.2016.18
17. 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). J Am Med Inform Assoc. 2010;17(2):124-30. DOI: 10.1136/jamia.2009.000893
18. Athey BD, Braxenthaler M, Haas M, Guo Y. tranSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. AMIA Jt Summits Transl Sci Proc. 2013 Mar 18;2013:6-8.

Corresponding author:
Christoph Beger
Universität Leipzig, Institut für Medizinische Informatik, Statistik und Epidemiologie, Härtelstraße 16–18, 04107 Leipzig, Germany
christoph.beger@imise.uni-leipzig.de

Please cite as
Beger C, Matthies F, Schäfermeier R, Uciteli A. Model-driven execution of phenotype algorithms – introduction of the Terminology- and Ontology-based Phenotyping Framework. GMS Med Inform Biom Epidemiol. 2023;19:Doc17.
DOI: 10.3205/mibe000256, URN: urn:nbn:de:0183-mibe0002565

This article is freely available from
https://doi.org/10.3205/mibe000256

Published: *2023-11-20*