Patient Cohort Identification on Time Series Data Using the OMOP Common Data Model

Christian Maier[1] Lorenz A. Kapsner[2] Sebastian Mate[2] Hans-Ulrich Prokosch[1,2] Stefan Kraus[3]

[1] Chair of Medical Informatics, Friedrich–Alexander–Universität Erlangen–Nürnberg (FAU), Erlangen, Bayern, Germany
[2] Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Erlangen, Bayern, Germany
[3] Department of Computer Science, Mannheim University of Applied Sciences, Mannheim, Baden-Württemberg, Germany

Address for correspondence Christian Maier, Chair of Medical Informatics, Friedrich–Alexander Universität Erlangen–Nürnberg, Am Wetterkreuz 15, 91058 Erlangen, Germany (e-mail: chris.maier@fau.de).

Appl Clin Inform 2021;12:57–64.

Abstract

Background The identification of patient cohorts for recruiting patients into clinical trials requires an evaluation of study-specific inclusion and exclusion criteria. These criteria are specified depending on corresponding clinical facts. Some of these facts may not be present in the clinical source systems and need to be calculated either in advance or at cohort query runtime (so-called feasibility query).

Objectives We use the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) as the repository for our clinical data. However, Atlas, the graphical user interface of OMOP, does not offer the functionality to perform calculations on facts data. Therefore, we were in search for a different approach. The objective of this study is to investigate whether the Arden Syntax can be used for feasibility queries on the OMOP CDM to enable on-the-fly calculations at query runtime, to eliminate the need to precalculate data elements that are involved with researchers' criteria specification.

Methods We implemented a service that reads the facts from the OMOP repository and provides it in a form which an Arden Syntax Medical Logic Module (MLM) can process. Then, we implemented an MLM that applies the eligibility criteria to every patient data set and outputs the list of eligible cases (i.e., performs the feasibility query).

Results The study resulted in an MLM-based feasibility query that identifies cases of overventilation as an example of how an on-the-fly calculation can be realized. The algorithm is split into two MLMs to provide the reusability of the approach.

Conclusion We found that MLMs are a suitable technology for feasibility queries on the OMOP CDM. Our method of performing on-the-fly calculations can be employed with any OMOP instance and without touching existing infrastructure like the Extract, Transform and Load pipeline. Therefore, we think that it is a well-suited method to perform on-the-fly calculations on OMOP.

Keywords

- information storage and retrieval
- clinical data
- electronic health records
- Arden Syntax
- Medical Logic Modules

received
June 22, 2020
accepted after revision
November 4, 2020

Georg Thieme Verlag KG,
Rüdigerstraße 14,
70469 Stuttgart, Germany

DOI https://doi.org/10.1055/s-0040-1721481.
ISSN 1869-0327.

Background and Significance

The identification of patient cohorts for recruiting patients into clinical trials requires an evaluation of study-specific inclusion and exclusion criteria.[1] Today, the availability of electronic health record data for secondary use and the opportunity to use software applications for this purpose relieves researchers from the laborious task of going through piles of paper records while looking for patients who match the study criteria.[2] A well-established research platform supporting feasibility queries and patient recruitment is Observational Health Data Sciences and Informatics (OHDSI), which builds on the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).[3] It has been proven as a suitable data model for storing medical data with standardized terminologies all around the world.[4–7] Furthermore, it includes several applications that enable the analysis of the patient data for different research purposes.[3] Hereby, Atlas is the central application that allows the researchers to specify eligibility criteria for patient recruitment by means of a web-based graphical user interface, enabling them to identify study-specific patient cohorts.[1]

Atlas is tailored toward modeling and executing feasibility queries that evaluate static facts data. However, some criteria may not initially be present as static facts, but need to be calculated first. There are two different approaches to perform such calculations. The common method is to precalculate additional facts during the so-called Extract, Transform and Load (ETL) process,[8] which constitutes the integration of patient data from clinical source systems into the OMOP CDM. A disadvantage of this method is that the ETL pipeline has to be adapted each time another to-calculate fact is asked for. An alternative method is the on-the-fly calculation of facts at query runtime, which is currently not natively supported by Atlas.

Our research, however, indicated a demand for on-the-fly calculations with respect to OHDSI/OMOP. A thread on the OHDSI forum addresses the problem of cohort selection on values that are not physically present as static facts in the source system and would therefore need to be calculated.[9] The proposed solution was to perform those calculations at ETL runtime. However, this method has some disadvantages, as we will discuss further in this article.

Ross et al found that 40% of the eligibility criteria in clinical trials relied on temporal data (i.e., time series data).[10] To make use of that kind of data for patient cohort identification, it necessitates involving a series of those static facts in a calculation. This underlines the need of (on-the-fly) calculations with regard to patient cohort identification.

At our local hospital, we observed a similar demand for on-the-fly calculations on temporal data when local researchers aimed to identify intensive care unit (ICU) patients who had been overventilated during an ICU stay.[11] The identification of overventilation is not possible with static facts alone, since it requires the calculation of a patient-individual critical limit and an aggregation of clinical time series data, that is, the expiratory tidal volume. A technology that proved suitable in an earlier investigation[12] for both types of calculations is the Arden Syntax standard.[13,14] Multiple previous investigations are available in scientific literature using it for patient cohort identification.[12,15–17] This suggests that the Arden Syntax might be a suitable technology for such on-the-fly calculations.

Therefore, the objective of this study is to investigate how feasibility queries on the OMOP CDM that involve calculations at query runtime can be realized with the Arden Syntax, to eliminate the need for precalculating data elements during the ETL process.

Methods

The Arden Syntax and PLAIN

The Arden Syntax is a Health Level Seven standard for clinical decision support (CDS) functions in the form of Medical Logic Modules (MLMs). An MLM corresponds to a condition-action rule, whereas the actual decision logic is implemented by means of programming language constructs in the form of statements and operators.[18] As the technical platform for our study, we applied the Arden Syntax processor used in clinical routine at our local hospital. It is based on a generalized Arden Syntax version termed PLAIN,[19] which is mostly compliant with the Arden Syntax version 2.8, but provides multiple extended features. The features relevant for this study are as follows: First, PLAIN provides native support of Logical Observation Identifiers Names and Codes (LOINC)[20] annotated data. Second, it provides a natively integrated data mapper for an XML format termed PLAIN Data Markup Language (PDML),[21] and third, it provides a shortcut for the call of other MLMs as user-defined functions (UDFs). These extended features will be explained in the examples below.

The Basic Approach

We conceived an approach that can be paraphrased as "looping through a sequence of records." In the context of our study, the term "record" refers to a data structure that includes all patient data necessary to evaluate the eligibility criteria. Such a record is a subset of the actual electronic medical record (EMR). Our basic approach is intentionally oriented toward what a researcher would do for cohort identification with paper-based records, that is, to consider one record after another, and to apply the eligibility criteria to each particular record.

As introduced in the Background and Significance section, our approach aims at identifying patients that were overventilated during an ICU stay. For that use case, it is important to note that a patient may have multiple ICU stays, either during a single hospital stay or a specific period of time. Consequently, we integrated the ICU case identity (ID) into the record and denoted it a "case record." This will allow determining during which ICU stays the patient has been overventilated. ►Fig. 1 (right) shows the basic structure of such a case record. It includes a case ID, a person ID, and an arbitrary collection of LOINC annotated medical facts.

►Fig. 1 (left) shows the PDML representation of a case record. It is based on an alternative approach for accessing EMRs from MLMs.[21] We slightly adapted this approach for our needs of accessing data sets from a CDM: Our case record

```
<object>
  <attribute name="caseID">
        <string>24931</string>
  </attribute>
  <attribute name="personID">
        <string>1036534</string>
  </attribute>
  <attribute name="data">
    <list>
      <string loinc="46098-0">LA2-8</string>
      <number loinc="3137-7">165</number>
      <number loinc="76007-4" time="2018-03-23 07:40:00">468</number>
      <number loinc="76007-4" time="2018-03-23 08:54:00">376</number>
      <number loinc="76007-4" time="2018-03-23 10:17:00">363</number>
      <number loinc="76007-4" time="2018-03-23 11:20:00">350</number>
      <number loinc="76007-4" time="2018-03-23 11:53:00">375</number>
      <number loinc="76007-4" time="2018-03-23 12:58:00">320</number>
      <number loinc="76007-4" time="2018-03-23 13:53:00">411</number>
      ......
    </list>
  </attribute>
</object>
```

case record	
caseID	string
personID	string
data	list of LOINC annotated facts

Fig. 1 Basic structure of a case record and its PLAIN Data Markup Language (PDML) representation.

is an equivalent to the CDS EMR described there but has been modified in a way to represent case-based data. Therefore, in our use case, a case record constitutes an object the attributes of which correspond to the basic structure in ►Fig. 1 (right). Its data section includes the patient's height (LOINC 3137–7), gender (46098–0), and the time series of the tidal volumes (76007–4).

Implementation of the Case Record Service

The OHDSI tools include a REST service called the Web API (https://github.com/OHDSI/WebAPI). It provides functions and data from the OMOP database to OHDSI applications like Atlas. Its function set also includes the cohort identification algorithm so that it is not tailored toward providing comprehensive data to applications for performing the cohort identification externally.

Therefore, we could not use the Web API for the retrieval of the OMOP data. Instead, we implemented a "case record service" with Talend Open Studio for Enterprise Service Bus[22] that fulfills the requirements of our approach. This case record service accesses the OMOP database and provides data in the form of a list of case records. It can be called via an URL and a list of parameters in the query string to define the selection criteria. The selection criteria allow passing a start and end date to define the period of time during which the patient visit should have taken place, as well as an optional array of LOINC codes to compose the contents of the data section of the case records. By means of the list of LOINC codes, the case record service filters out the ICU cases in which not every corresponding LOINC code was documented at least one time.

►Fig. 2 illustrates the process of retrieving data from the case record service. The first step involves the HTTP GET request (with parameters defined in the query string) to the case record service. The case record service processes the parameters and constructs SQL queries that are performed on the OMOP database in step 2. For our use case, the necessary data are read from the measurement and visit_occurrence tables of OMOP and returned back to the case

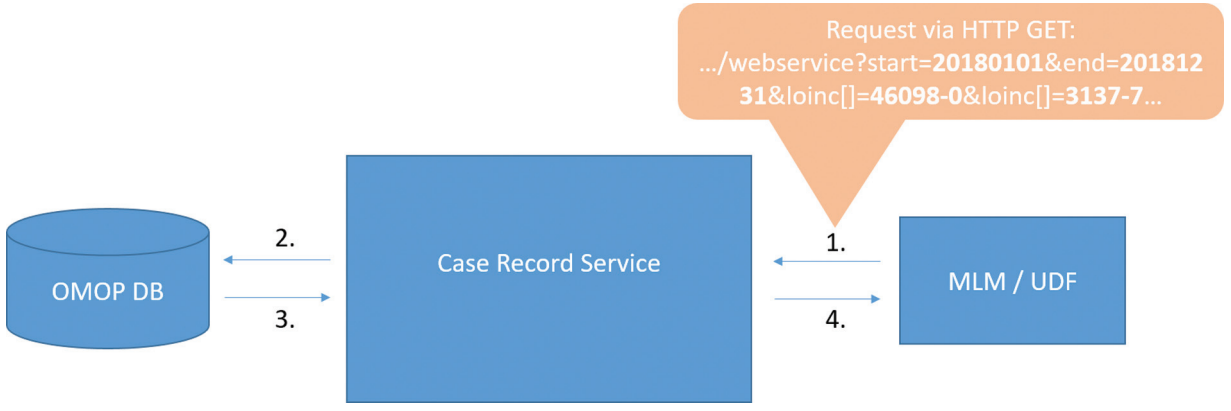

Fig. 2 The process of retrieving data from the case record service.

record service in step 3. Finally, the case record service generates the PDML representation and returns it to the MLM in step 4, which can now perform the on-the-fly calculation.

Implementation of the Feasibility Query

We implemented an MLM that we denote as the "query MLM." It retrieves a list of case records from the case record service. Then, it loops through this list and applies the eligibility criteria to each case record.

The identification of overventilated patients requires the calculation of the patient-individual critical limit for the expiratory tidal volume. Therefore, we implemented the UDF "tv_critical" that is called from the query MLM for each record. The UDF is passed the case record as a parameter so that all necessary data for the on-the-fly calculation of the before-said critical limit is available in the UDF. Then, it extracts the LOINC annotated tidal volumes from the data section of the particular case record, calculates the patient-individual critical limit, and returns it to the query MLM. Finally, the query MLM identifies a possible overventilation depending on that information.

To verify our approach, we additionally implemented the same query in R and compared the resulting patient set with the one identified by the query MLM/UDF "tv_critical." We used the R package "dplyr,"[23] which is also commonly used with its special syntax in several projects of the OMOP Methods Library (https://www.ohdsi.org/methods-library/).

Results

Our study resulted in an MLM-based feasibility query that identifies cases of overventilation. The query employs two different MLMs. The first MLM is the actual query MLM that processes a sequence of case records and successively applies the eligibility criteria. The second MLM is the UDF "tv_critical," (►Fig. 4) which calculates the patient-individual critical limit for the tidal volume and is called from the query MLM.

The Query MLM

The query MLM is shown in ►Fig. 3. Line 1 retrieves the list of case records from the case record service. Line 3 initializes an empty list of eligible cases. The centerpiece of the query MLM is a for loop, starting in line 5, that iterates over the case record list. Line 7 extracts the tidal volumes from the particular case record by means of the EXTRACT LOINC statement, using the corresponding LOINC code 76007–4. The conditional statement, starting in line 9, applies the eligibility criteria, which verifies whether the average of the tidal volumes is greater than the patient-individual limit calculated by the UDF "tv_critical." If the criterion is fulfilled, the case ID is added to the result list eligible.

The UDF

To provide the patient-individual critical limit for the tidal volume, the UDF "tv_critical" calculates the predicted body weight (line 6–10), based on the patient's gender (LOINC 46098–0; LOINC answer ID LA2–8 for male and LA3–6 for female patients) and height (LOINC 3137–7). When the query MLM calls "tv_critical," it passes the particular case record as the argument.

Verification and Comparison of Runtime

The comparative implementation of our approach showed that both the R script and the MLM-based approach identify exactly the same cases. Furthermore, we measured the calculation runtime of both approaches. The underlying data set comprises 75,777 LOINC annotated tidal volumes in total and a mean of 57.36 tidal volumes per clinical case.

The runtime of the MLM-based approach took 1,020 ms whereas the R script took 350 ms. The main reason for the difference in runtime is the input format that is different for each solution. The MLM is passed the PDML representation of the case records (►Fig. 1), whereas the R script uses a comma-separated values (CSV) format that has been generated from the PDML due to easier handling of CSV in R. For the Arden implementation, it is also possible to catch up with the

```
1.  recordlist := READ "http://intranet.uk-
    erlangen.de/omop/webservice?start=20180101&end=20181231&loinc[]=46098-
    0&loinc[]=3137-7&loinc[]=76007-4";
2.
3.  eligible := [];
4.
5.  FOR EACH record IN recordlist DO
6.
7.       tidalvolumes := EXTRACT LOINC "76007-4" FROM record;
8.
9.       IF THE AVERAGE OF tidalvolumes IS GREATER THAN (@tv_critical record) THEN
10.          ADD record.caseid TO eligible;
11.      ENDIF;
12.
13. ENDDO;
14.
15. WRITE THE COUNT OF eligible;
```

Fig. 3 Query Medical Logic Module (MLM) that loops through the case records retrieved from the case record service and calls the user-defined function (UDF) "tv_critical."

```
record := ARGUMENT;

gender := EXTRACT SINGLE LOINC "46098-0" FROM record;
height := EXTRACT SINGLE LOINC "3137-7" FROM record;

IF gender IS EQUAL "LA2-8" THEN
  predicted_body_weight := 50 + 0.91 * (height - 152.4);
ELSEIF gender IS EQUAL "LA3-6" THEN
  predicted_body_weight := 45.5 + 0.91 * (height - 152.4);
ENDIF;

RETURN predicted_body_weight * 8.8;
```

Fig. 4 User-defined function (UDF) "tv_critical" that calculates the predicted body weight and the patient-specific critical tidal volume and returns it to the query Medical Logic Module (MLM).

runtime of the R script when the PDML is preprocessed first (~ 155 ms runtime). However, we think that the actual difference in calculation times of the standard PDML representation will not be a limiting factor for a potential practical use.

For the interested reader, we provide the R scripts in the ►**Supplementary Material** (available in the online version).

Discussion

Our study constitutes a proof of concept for MLM-based feasibility queries on the OMOP CDM involving on-the-fly calculations. The conventional way would be to run those calculations in advance during the ETL process, which would eventually result in additional static facts in the OMOP CDM. However, this requires (1) additional computational performance for the data transformation process, (2) leads to an increase of complexity of the ETL pipeline, and finally, (3) it is impossible to foresee calculated facts required by researchers in the future. Moreover, researchers rarely have access to the ETL source code (such as SQL queries) and are usually not trained to modify it, so that there would be the need to involve a computer scientist. In contrast, the on-the-fly approach presented in this article does not require modifying the ETL pipeline.

In terms of execution time, it is difficult to compare an ETL-based approach with an on-the-fly calculation as described in this study. ETL implementations are proprietary pipelines and differ depending on the underlying source data and the implementation environment used. Due to these circumstances, we do not think that it is generally valid to make a comparison of a proprietary ETL pipeline that extracts data from medical records to the case record service querying an OMOP instance.

Identification of Cases versus Patients

Typically, research query tools aim to identify cohorts on the patient level. While Atlas provides additional (graphical) statistics of the resulting patient cohort, other query tools, such as Informatics for Integrating Biology & the Bedside (i2b2),[24] are tailored to return "just" several patients that match the given research criteria. The cohort identification at patient level is sufficient for a huge number of clinical studies. However, for our use case of overventilation the granularity of the finest entity needs to be the ICU stay instead of the patient, as it can occur that a patient has been in intensive care multiple times. This is why our researchers were interested in identifying ICU cases instead of patients because a patient may be overventilated during a specific ICU stay, but not in another.

The Arden Syntax for Cohort Identification

The Arden Syntax standard was not originally designed for performing MLM-based feasibility queries. Nevertheless, the literature shows repeated attempts to use it for this purpose. In a review on the formal representation of eligibility criteria, Weng et al describe the Arden Syntax as a suitable technology with respect to the expressiveness of the language constructs.[15] Ohno-Machado et al used a modified Arden Syntax version for the identification of eligible patients for a breast cancer trial[17]; the modifications to the standard to facilitate patient recruitment have been described by Wang et al.[16] Sarkar et al described the use of MLMs to identify very low birth weight neonates.[25] In 2017, Mate et al demonstrated the use of MLMs to postprocess i2b2 queries,[12] based on the standard compliant Arden Syntax environment Arden2Bytecode.[26]

From a mere technical perspective, on-the-fly calculations for cohort identification can be performed with any common programming language. However, we believe that the Arden Syntax has some advantages with regard to comprehensibility that we discuss further in the "Generalizability of the Method" section.

Generalizability of the Method

Our approach is based on an experimental Arden Syntax processor. Transferring it to standard-compliant Arden Syntax environments takes some extra steps. First, conventional Arden Syntax MLMs require a frame-like structure that, in contrast to PLAIN MLMs, cannot be omitted. Second, PLAIN provides a natively integrated XML mapper, while in Arden Syntax it must first be integrated by a programmer. Third, the Arden Syntax standard does not natively support LOINC annotations. Consequently, the LOINC codes must be provided in the case records in the form of object attributes, complicating their structure and the code to process them. In conclusion, it is possible to use standard-compliant processors for our approach, but an adapted version is clearly more comprehensible.

The Web API implements server-side functions for cohort identification that can be called from external sources. It is intended to be the service that is used by Atlas to retrieve (aggregated) data from the OMOP repository. As it is a REST web service and returns the data in a standardized format (JSON, etc.), it can also be used by other applications except Atlas. This prompted us to investigate whether there is a function of the Web API that would fulfill the needs of our approach. Our literature review revealed a study in which the authors could successfully use the Web API for interaction with a proprietary tool: Yuan et al described the use of the Web API as a translation tool for external created cohort definitions.[27] We identified a function that could provide patient data for a specific patient from the measurement and observation table from OMOP, which is exactly what we provide for in the data section of a single case record. However, it does not offer to retrieve a bundle of patient data of multiple patients or cases. Therefore, the corresponding function would not allow retrieving a data set that would be sufficient for running feasibility queries on it. A study by Unberath et al describes the extension of the Web API to retrieve observation and measurement data for a single patient.[28] The adaption has been implemented, as the Web API did not offer a function to retrieve these data through its original function set, which is a similar problem that we encountered in our use case.

We considered whether we should also make an adaptation of the Web API, but finally decided to implement the proprietary case record service to be independent of any future changes to the Web API. Beyond that, an adaptation of the Web API is only sustainable if it is agreed with OHDSI and included in the official Web API trunk, which we considered as a longer lasting process without any guarantee. Thus, the case record service can be seen as an independent add-on to the existing OMOP/OHDSI ecosystem.

We believe that our approach of using query MLMs and UDFs to enable cohort identification involving on-the-fly calculations is transferable to other use cases. Any other query MLM would implement the same basic principle and therefore would strongly resemble ►Fig. 3. The URL of the case record service, the LOINC codes, and the condition of the IF statement would have to be modified. Moreover, other query MLMs will likely invoke other UDFs, depending on the specific eligibility criteria. The case record service is reusable for other use cases as it simply provides the data from the OMOP repository.

Our approach does not necessarily involve time series data, but can be used for calculations based on any combination of static facts that are not part of a series. An example is the calculation of the body mass index, which is an item that is rarely available as a precalculated value in the database of clinical source systems.

Comprehensibility of Our Approach

The designers of the Arden Syntax put great emphasis on easy comprehensibility of the language constructs, hoping that clinicians will be able to read and understand MLMs with little training.[13] This is also an important aspect with respect to query MLMs, since it would be of great benefit if researchers can understand and thus verify them. It might even be worth a consideration whether researchers with algorithmic skills are able to create query MLMs on their own. Of course, we cannot expect researchers to be familiar with Arden Syntax, and delegating the programming task to software engineering professionals may be the most reasonable solution, as Nadkarni suggests.[29] Nevertheless, Arden Syntax is a domain-specific language (DSL)[30] that some consider easier to learn than the common programming languages.[14] In his blog, Fowler outlines that understanding a DSL, even if one cannot write code, the ability to read and understand it "can build a deep and rich communication channel between software development and the underlying domain."[31] We agree with this opinion and believe that a simple language that researchers can at least understand would be of great practical advantage. In the field of CDS functions and expert systems, it is common knowledge that the lack of such a communication channel is problematic and leads to a knowledge acquisition bottleneck.[32]

Limitations and Future Work

Currently, our approach does not allow for the definition of other prefilter criteria beside a time range and an array of LOINC identifiers, so that it only constitutes a proof of concept. For future work, we plan to extend our approach by using the export of cohort definitions in Atlas. These are available as JSON or several other formats and represent the inclusion and exclusion criteria that were defined in the Atlas frontend before.[33] This cohort definition could then be sent to the case record service to prepare the patient data and would therefore make the prefilter function obsolete. This way, only the data that matches the criteria definitions will be returned to be processed further by the query MLM.

For the practical use of our approach, there is probably a need of realizing a more complex end-to-end system. We think that the before-said export of the cohort definitions in Atlas could be a first step. Another possibility would be the integration of an execution environment for Arden in Atlas. However, for both issues to be sustainable there would be the need to involve OHDSI to achieve a full integration of our solution with the existing OHDSI environment.

The Arden Syntax does not provide native support of terminologies such as LOINC. In contrast, PLAIN provides experimental support for LOINC, so that we think it is an appropriate solution for this study. However, other use cases might require the involvement of different terminologies. This would currently not be possible with PLAIN so that it would be worth a discussion to integrate native support for additional terminologies. OMOP requires storing medical facts linked to a medical terminology. Hereby, it does not matter which terminology is used as long as it is available as a vocabulary in OMOP. We decided to use LOINC as the terminology for our data representation as it comprehensively covers clinical and laboratory values, which are often subject to time series analyses. It also includes concepts for demographics, procedures, surveys, and corresponding results. Furthermore, it is free to use and PLAIN offers easy access via its LOINC SEARCH feature.

Conclusion and Outlook

We found that MLMs are a suitable technology for feasibility queries on the OMOP CDM. In conjunction with the case record service, our method of performing on-the-fly calculations can be employed with any OMOP instance and, more importantly, without touching any existing infrastructure like the ETL pipeline. Therefore, we think that it is a straightforward approach to perform on-the-fly calculations on OMOP.

Clinical Relevance Statement

Researchers face challenges when using Atlas for cohort identification purposes. Criteria evaluation may refer to patient data that are not available as static facts, and thus require calculations, either during ETL or at runtime. Researchers usually cannot extend the database on their own. Consequently, a means of enabling on-the-fly calculations is practically relevant.

Multiple Choice Questions

1. Why did the described approach require an implementation of the case record service instead of using the Web API for the data retrieval?
   a. The Web API only allows the definition of criteria instead of returning data.
   b. The Web API does not implement a function that would return a data set that is sufficient for running feasibility queries on it.
   c. The Web API can only be used as service by Atlas.
   d. The format of the data that is returned by the Web API cannot be processed by proprietary tools.

   Correct Answer: The correct answer is option b.

2. What is the reason for choosing a generalized Arden Syntax version for this approach instead of the standard Arden Syntax?
   a. PLAIN offers several language constructs that proved beneficial for the methods described.
   b. Only PLAIN can be used for patient cohort identification.
   c. The standard Arden Syntax is outdated and should not be used anymore nowadays.
   d. The standard Arden Syntax cannot process LOINC annotated data sets, even with workarounds.

Correct Answer: The correct answer is option a.

Protection of Human and Animal Subjects

Only anonymized data was used. Therefore, the authors declare that the study was conducted in accordance with the ethical principles of the Helsinki Declaration.

Supplementary Material

The supplementary material contains the R scripts that serve for the verification of the patients identified by the Arden implementation.

Funding

This work was funded in part by the German Federal Ministry of Education and Research (BMBF) within the Medical Informatics Initiative (MIRACUM Consortium) under the Funding Number FKZ: 01ZZ1801A. The present work was performed in fulfillment of the requirements for obtaining the degree “Dr. rer. biol. hum.” from the Friedrich-Alexander-Universität Erlangen-Nürnberg (CM).

Conflict of Interest

None declared.

References

1 Meystre SM, Heider PM, Kim Y, Aruch DB, Britten CD. Automatic trial eligibility surveillance based on unstructured clinical data. Int J Med Inform 2019;129:13–19
2 Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit Translat Bioinforma 2010;2010:1–5
3 Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574–578
4 Maier C, Lang L, Storf H, et al. Towards implementation of OMOP in a German University Hospital Consortium. Appl Clin Inform 2018; 9(01):54–61
5 Lamer A, Depas N, Doutreligne M, et al. Transforming French Electronic Health Records into the Observational Medical Outcome Partnership's Common Data Model: a feasibility study. Appl Clin Inform 2020;11(01):13–22
6 Yoon D, Ahn EK, Park MY, et al. Conversion and data quality assessment of electronic health record data at a Korean tertiary teaching hospital to a Common Data Model for distributed network research. Healthc Inform Res 2016;22(01):54–58
7 Lynch KE, Deppen SA, DuVall SL, et al. Incrementally transforming electronic medical records into the Observational Medical Outcomes Partnership Common Data Model: a multidimensional quality assurance approach. Appl Clin Inform 2019;10(05):794–803
8 Denney MJ, Long DM, Armistead MG, Anderson JL, Conway BN. Validating the extract, transform, load process used to populate a large clinical research database. Int J Med Inform 2016;94:271–274
9 Defining a cohort in Atlas through a ratio between two measurement values OHDSI Forum. Published January 9, 2020. Accessed January 9, 2020 at: https://forums.ohdsi.org/t/defining-a-cohort-in-atlas-through-a-ratio-between-two-measurement-values/6768.
10 Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. Summit On Translat Bioinforma 2010;2010:46–50
11 Castellanos I, Martin M, Kraus S, et al. Effects of staff training and electronic event monitoring on long-term adherence to lung-protective ventilation recommendations. J Crit Care 2018;43:13–20
12 Mate S, Castellanos I, Ganslandt T, Prokosch H-U, Kraus S. Standards-based procedural phenotyping: the Arden Syntax on i2b2. Stud Health Technol Inform 2017;243:37–41
13 Hripcsak G, Ludemann P, Pryor TA, Wigertz OB, Clayton PD. Rationale for the Arden Syntax. Comput Biomed Res 1994;27(04):291–324
14 Samwald M, Fehre K, de Bruin J, Adlassnig K-P. The Arden Syntax standard for clinical decision support: experiences and directions. J Biomed Inform 2012;45(04):711–718
15 Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. J Biomed Inform 2010;43(03):451–467
16 Wang SJ, Ohno-Machado L, Mar P, Boxwala AA, Greenes RA. Enhancing Arden Syntax for clinical trial eligibility criteria. Proc AMIA Symp . Published online 19991188
17 Ohno-Machado L, Wang SJ, Mar P, Boxwala AA. Decision support for clinical trial eligibility determination in breast cancer. Proc AMIA Symp . Published online 1999340–344

18 Hripcsak G. Writing Arden Syntax Medical Logic Modules. Comput Biol Med 1994;24(05):331–363
19 Kraus S. Generalizing the Arden Syntax to a common clinical application language. Stud Health Technol Inform 2018;247:675–679
20 McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003;49(04):624–633
21 Kraus S, Toddenroth D, Staudigel M, et al. Mapping the entire record-an alternative approach to data access from Medical Logic Modules. Appl Clin Inform 2020;11(02):342–349
22 Talend Inc Open Source ESB: Talend Open Studio Free ESB Tool. https://www.talend.com/products/application-integration/esb-open-studio/. Accessed January 24, 2020
23 Wickham H, François R, Henry L, Müller K. dplyr: a grammar of data manipulation. https://dplyr.tidyverse.org. Published online 2019
24 Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(02):124–130
25 Sarkar IN, Chen ES, Rosenau PT, Storer MB, Anderson B, Horbar JD. Using Arden Syntax to identify registry-eligible very low birth weight neonates from the electronic health record. AMIA Annu Symp Proc 2014;2014:1028–1036
26 Gietzelt M, Goltz U, Grunwald D, et al. ARDEN2BYTECODE: a one-pass Arden Syntax compiler for service-oriented decision support systems based on the OSGi platform. Comput Methods Prog Biomed 2012;106(02):114–125
27 Yuan C, Ryan PB, Ta C, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. J Am Med Inform Assoc 2019;26(04):294–305
28 Unberath P, Prokosch HU, Gründner J, Erpenbeck M, Maier C, Christoph J. EHR-independent predictive decision support architecture based on OMOP. Appl Clin Inform 2020;11(03): 399–404
29 Nadkarni PM. Metadata-Driven Software Systems in Biomedicine: Designing Systems That Can Adapt to Changing Knowledge. London: Springer; 2011
30 Fowler M, Parsons R. Domain-Specific Languages. Boston: Addison-Wesley; 2011
31 Fowler M. Business Readable DSL. Accessed January 31, 2020 at: https://martinfowler.com/bliki/BusinessReadableDSL.html.
32 Sonntag D, Wennerberg P, Buitelaar P, Zillner S. Pillars of ontology treatment in the medical domain. In: Cases on Semantic Interoperability for Information Systems Integration: Practices and Applications IGI Global Hershey. 2010:162–186
33 Evans CC, Simonov K. Query combinators: domain specific query languages for medical research. Bioinformatics 2019. Doi: 10.1101/737619