German Medical Data Sciences: Visions and Bridges
R. Röhrig et al. (Eds.)

© 2017 German Association for Medical Informatics, Biometry and Epidemiology (gmds) e.V. and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms
of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/978-1-61499-808-2-37

Standards-Based Procedural Phenotyping: The Arden Syntax on i2b2

Sebastian MATE [a,1], Ixchel CASTELLANOS [b], Thomas GANSLANDT [c], Hans-Ulrich PROKOSCH [a,c] and Stefan KRAUS [a]
[a] *Medical Informatics, Univ. of Erlangen-Nürnberg, Erlangen, Germany*
[b] *Department of Anesthesiology, University Hospital Erlangen, Erlangen, Germany*
[c] *Center for Medical Information and Communication, University Hospital Erlangen, Erlangen, Germany*

Abstract. Phenotyping, or the identification of patient cohorts, is a recurring challenge in medical informatics. While there are open source tools such as i2b2 that address this problem by providing user-friendly querying interfaces, these platforms lack semantic expressiveness to model complex phenotyping algorithms. The Arden Syntax provides procedural programming language construct, designed specifically for medical decision support and knowledge transfer. In this work, we investigate how language constructs of the Arden Syntax can be used for generic phenotyping. We implemented a prototypical tool to integrate i2b2 with an open source Arden execution environment. To demonstrate the applicability of our approach, we used the tool together with an Arden-based phenotyping algorithm to derive statistics about ICU-acquired hypernatremia. Finally, we discuss how the combination of i2b2's user-friendly cohort pre-selection and Arden's procedural expressiveness could benefit phenotyping.

Keywords. phenotyping, cohort identification, i2b2, Arden Syntax

1. Introduction

The process of identifying patient cohorts based on a set of characterizing patient data elements is commonly referred to as "phenotyping" in the literature. According to the SHARPn project [1], phenotyping is the algorithmic recognition of patients. It includes clinical trials, quality metrics, outcomes research, observational studies, decision support, and other tasks [1]. The search strategy to identify a patient cohort is called "phenotyping algorithm". According to [1], such algorithms consist of complex sets of inclusion and exclusion criteria, coupled using sets of logical operators. Mo et al. [2] have found that "phenotyping algorithms can involve multiple complex logical steps, integrating various operations". These can be Boolean logic, numeric comparator operations, arithmetic functions (e.g. to calculate BMI), aggregative operations (e.g. COUNT, FIRST), negation (negative assertion vs. exclusion/empty set) and temporal relations between events [2]. Ross et al. [3] found that 85% of the eligibility criteria of clinical trials had a significant semantic complexity and 40% relied on temporal data.

The widely used *Informatics for Integrating Biology and the Bedside* (i2b2) [4,5] is a software platform that can be used for basic phenotyping. The system contains

[1] Corresponding Author: **Sebastian.Mate@fau.de**

special features such as automatic unit conversion or conceptual subsumption to automatically query for child concepts as found in hierarchically structured terminologies. The query logic relies on Boolean logic, numeric comparison operations, negation and temporal aspects. However, as outlined by Mo et al. [2], the complexity of phenotyping algorithms (like those in eMERGE [6]) can exceed the capabilities of platforms such as i2b2. While the current i2b2 version 1.7 has included support for complex temporal relations, there are still some limitations. For example, i2b2 limits numeric comparators to fixed values (e.g. "$A > 120$") and it is not possible to model relative comparisons (e.g. "$A > B$"). Furthermore, it does not support arithmetic functions, which could be used to compute criteria on-the-fly (e.g. "$BMI = weight / height^2$").

The Arden Syntax for Medical Logic Systems [7] may be a suitable technology to address these shortcomings. It is an HL7 standard that was designed to enable clinical decision support functions in the form of Medical Logic Modules (MLMs), which are typically used to monitor clinical events. It was intended to be user-friendly, enabling even non-experts in computer science to write MLMs. It also features a rich set of operators and a time-stamped data type system that is tailored to medical data.

In this paper, we investigate if i2b2 can be used together with the Arden Syntax to go beyond the functionality of current phenotyping systems. It is our aim to provide a proof of concept and to discuss the assets and drawbacks of this approach.

2. Methods

We created a prototypical Java tool. It contains an integrated MLM editor that uses the RSyntaxTextArea library (https://github.com/bobbylight/RSyntaxTextArea) for which we implemented Arden Syntax highlighting (see Figure 1).

We then developed and integrated a method for translating i2b2 queries into "template" MLMs. By deriving and parsing the i2b2 query XML definitions from the i2b2 database, we create Arden Syntax lists for all i2b2 ontology concepts that were used in the original query. To populate these with data, our tool creates Arden "curly braces" expressions, which contain the SQL statements to retrieve the data records directly from the i2b2 database. It then adds further Arden code, which merges all lists into one central data structure in the form of a list of patient objects, which we called "PatientData". Each patient object within "PatientData" provides an attribute for the patient number, as well as one attribute for each original i2b2 query concept. To assist the user with later converting the "template" MLM into a true phenotyping algorithm, a comment is automatically generated and inserted into the MLM to reveal the available attributes of the "PatientData" object (Figure 1, first visible line of code).

Finally, we integrated Arden2ByteCode, a Java-based, open source Arden Syntax environment [8], into our tool. The execution of an MLM in Arden2ByteCode can be triggered from within our program, which also displays the result of the execution.

3. Results

Our implementation allows for post-processing i2b2 query results with Arden Syntax MLMs by applying additional (more restrictive) filtering or computations on patient data, which have been found with the i2b2 system previously. The user has to go through the following steps to use our system:

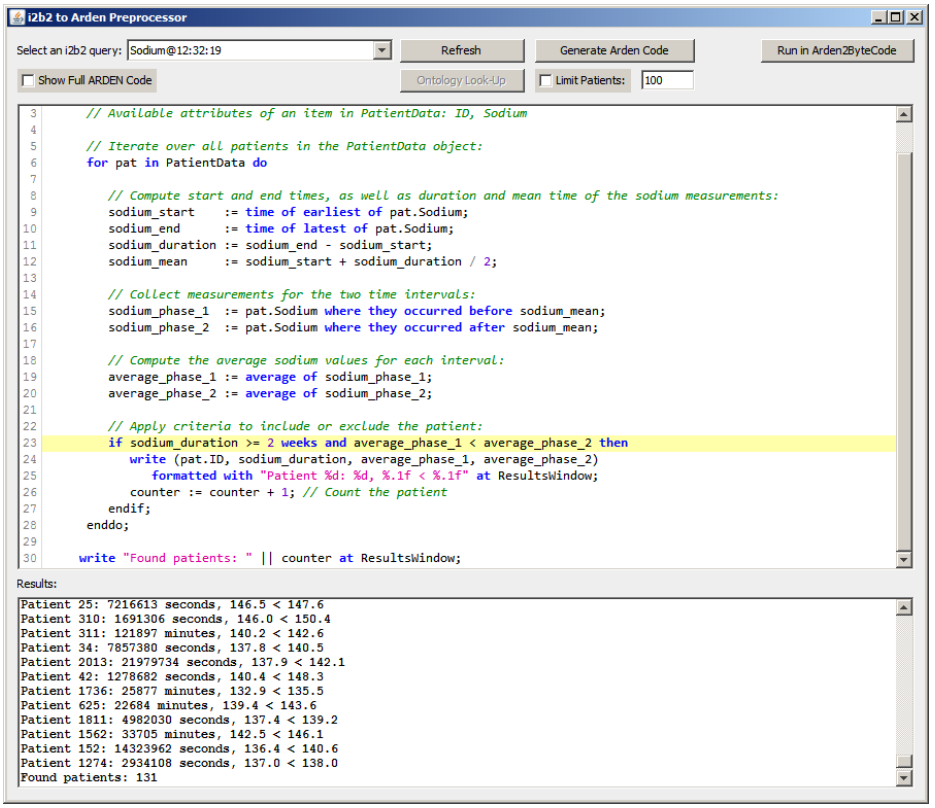

Figure 1. i2b2Arden user interface with a phenotyping algorithm (patient numbers anonymized).

1. The user creates and executes a query within the i2b2 environment. This query includes the clinical data elements of interest, along with their Boolean relations. It may also use other i2b2 features, such as temporal relations and value restrictions.
2. The user opens our tool and selects this i2b2 query. The tool then analyzes the i2b2 query and automatically prepares an MLM template. This transforms the facts data of the i2b2 query's data elements into appropriate Arden Syntax data structures.
3. The user adds further phenotyping logic, based on the Arden Syntax' rich set of around 150 operators, to the MLM. This may include complex temporal logic, operations on values and comparisons between multiple data elements.

To test our prototype in a real environment, we designed and performed a data analysis for the largest 35-bed ICU at our local University Hospital. ICU-acquired hypernatremia (IAH) is a commonly described phenomenon, where sodium blood values increase during hospitalization. Risk factors associated with IAH have been reported, such as male gender and age above 50 [9].

We exported the patient age, gender and sodium measurements since 2015 from our ICU system and uploaded these into an i2b2 project on our i2b2 version 1.7.07 instance. As described above, the first step was to run a query in i2b2. In our example, this initially included only one concept, "*Sodium*". This i2b2 query returned 2,242 patients (the whole database). After creating the MLM template with our tool, we cre-

ated the phenotyping algorithm, which is shown in Figure 1. The program counts and lists all applicable patients with increasing sodium values. By replacing the "<" comparator in line 23 with ">", the patients with decreasing sodium were found. For 43.8%, the sodium values were increasing. We then repeated the test and modified the initial i2b2 query to include the constraints "*Age > 50 years*" and "*Gender = Male*". After running the Arden Syntax MLM again, we found 48.5% of second group having increasing values, which supports the findings of [9].

4. Discussion

While our phenotyping MLM still needs refinement and further evaluation of the results, it demonstrates that our approach is capable of executing complex procedural phenotyping algorithms, and is able to benefit from the easy-to-use, graphical preselection of patient cohorts within i2b2. We plan further investigations on how to design easy-to-use, yet powerful, phenotyping environments. Tools such as i2b2 were not intended to be fully featured data analysis or dedicated phenotyping environments, but rather meant to serve as user-friendly hypothesis generation and validation tools to allow for pre-selecting patient cohorts. Yet, it would still be useful for researchers not having to fall back on low-level technology (such as SQL) or complex statistical software for pursuing complex phenotyping. It is a matter of discussion where to draw the line between user-friendliness and computational power. Future work might focus on integrating our Arden code editor directly into the i2b2 workbench.

Similar approaches are described in the literature. For example, there are various "R" environment integrations for i2b2 (e.g. [10]). These, however, do not allow for "live" editing of program code, in contrast to our approach. There are also other procedural approaches (e.g. [11]), but these do not build upon i2b2 or similar platforms to enable an easy-to-use pre-selection of patient cohorts.

From a technical point of view, computationally equivalent "phenotyping power" could be achieved in any Turing-complete programming language. However, we believe that our example illustrates that the Arden Syntax code might be easier to understand for non-experts in computer science, which was one of the design goals of the Arden Syntax standard [12]. Applying an all-purpose programming language instead may considerably reduce the number of potential users.

Our tool is of prototypical character and there is room for future improvements. It is based on Arden Syntax version 2.5, which is the latest version supported by Arden2Bytecode. Therefore we had to make some concessions: As described above, our tool prepares an MLM to include all concepts from the i2b2 query. This has been implemented by embedding SQL queries to access the i2b2 database. However, as they only return a database record set, we had to post-process these data records to properly align values and their associated time stamps. Otherwise it would not be possible to make use of certain Arden Syntax constructs, such as *time of*. This post-processing requires about 40 lines of additional Arden Syntax code for each concept that has been used in the i2b2 query. Our current workaround is to allow hiding the automatically generated code in our tool via a checkbox (as shown in Figure 1). Therefore support for later versions of the Arden Syntax would considerably facilitate the integration process. In particular, the *as time* operator introduced in version 2.8 would save the additional code required to transform the string representation of a timestamp into the time data type of the Arden Syntax.

Finally, another aspect of later versions of the Arden Syntax is the support for fuzzy logic. The possibility of applying "soft" inclusion and exclusion criteria instead of "hard" Boolean logic could improve current patient identification methods.

5. Acknowledgements

The present work was performed in (partial) fulfillment of the requirements for obtaining the degree "Dr. rer. biol. hum." from the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (SM).

6. Human Subjects Protection

This study is in accordance with the German Bavarian Hospital Law (BayKrg §27). Only anonymized routine care patient data were used, no formal intervention was performed and no additional patient data were collected.

7. Software Availability

The source code and the full demonstration MLM from this paper are available on GitHub (https://github.com/sebmate/i2b2Arden).

References

[1] Rea S, Pathak J, Savova G, *et al.,* Building a Robust, Scalable and Standards-Driven Infrastructure for Secondary Use of EHR Data: The SHARPn Project, *J Biomed Inform* **45** (2012), 763–71.

[2] Mo H, Thompson WK, Rasmussen LV, *et al.,* Desiderata for Computable Representations of Electronic Health Records-Driven Phenotype Algorithms, *J Am Med Inform Assoc* **22** (2015), 1220–30.

[3] Ross J, Tu S, Carini S, *et al.,* Analysis of Eligibility Criteria Complexity in Clinical Trials, *AMIA Summits Transl Sci Proc* (2010), 46–50.

[4] Kohane IS, Churchill SE, Murphy SN, A Translational Engine at the National Scale: Informatics for Integrating Biology and the Bedside, *J Am Med Inform Assoc* **19** (2012), 181–5.

[5] Murphy SN, Weber GM, Mendis ME, *et al.,* Serving the Enterprise and Beyond with Informatics for Integrating Biology and the Bedside (i2b2), *J Am Med Inform Assoc* **17** (2012), 124–30.

[6] McCarty CA, Chisholm RL, Chute CG, *et al.,* The eMERGE Network: A Consortium of Biorepositories Linked to Electronic Medical Records Data for Conducting Genomic Studies, *BMC Med Genomics* **4** (2011), 13.

[7] Pryor TA, Hripcsak G, The Arden Syntax for Medical Logic Modules, *Int J Clin Monit Comput* **10** (1993), 215-224.

[8] 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 Programs Biomed* **106** (2012), 114–25.

[9] Alansari M, Abdulmomen A, Hussein M, *et al.,* Acquired Hypernatremia in a General Surgical Intensive Care Unit: Incidence and Prognosis, *Saudi J Anaesth* **10** (2016), 409.

[10] Segagni D, Ferrazzi F, Larizza C, *et al.,* R Engine Cell: Integrating R Into the i2b2 Software Infrastructure, *J Am Med Inform Assoc* **18** (2011) 314–7.

[11] Li D, Endle CM, Murthy S, *et al.,* Modeling and Executing Electronic Health Records Driven Phenotyping Algorithms Using the NQF Quality Data Model and JBoss® Drools Engine, *AMIA Annu Symp Proc* (2012), 532–41.

[12] Samwald M, Fehre K, de Bruin J, *et al.,* The Arden Syntax Standard for Clinical Decision Support: Experiences and Directions, *J Biomed Infor* **45** (2012), 711–8.