

Improving the Implementation of Clinical Decision Support Systems

Stefan Rüping¹, Alberto Anguita², Anca Bucur³, Traian Cristian Cirstea⁴, Björn Jacobs⁵, and Antje Torge⁶

Abstract—Clinical decision support (CDS) systems promise to improve the quality of clinical care by helping physicians to make better, more informed decisions efficiently. However, the design and testing of CDS systems for practical medical use is cumbersome. It has been recognized that this may easily lead to a problematic mismatch between the developers' idea of the system and requirements from clinical practice. In this paper, we will present an approach to reduce the complexity of constructing a CDS system. The approach is based on an ontological annotation of data resources, which improves standardization and the semantic processing of data. This, in turn, allows to use data mining tools to automatically create hypotheses for CDS models, which reduces the manual workload in the creation of a new model. The approach is implemented in the context of EU research project *p-medicine*. A proof of concept implementation on data from an existing Leukemia study is presented.

I. INTRODUCTION

Medical knowledge is growing at an unprecedented rate. New evidence is found and gradually brought into clinical care, new medication and treatments are introduced, and evidence-based guidelines are evolving and expanding.

Clinical decision support (CDS) systems promise to improve the quality of clinical care by helping the physicians to efficiently access and apply all the relevant clinical knowledge in order to make better decisions that are consistent with the latest developments in the medical state of the art. A clinical decision support system has the role to assist the clinical users in assessing the disease status of a patient, in making a diagnosis, in selecting a suitable treatment or in making other clinical decisions relevant for a patient case.

CDS tools need to be developed and strictly evaluated and validated before they can be integrated in clinical daily care or prospective clinical trials for decisionmaking. This, however, is a very complex task. For example, it has been reported that the CDS for respiratory care at the LDS

Hospital, Salt Lake City, USA – a well known example of a CDS system – has been developed with an effort of about 25 person years [1].

As a result, decision support systems are not yet systematically used. It has been identified that a major reason for this lack of a practical uptake is a mismatch between the developers' and the clinicians' views, i.e. that many attempts of developing such systems are *bound to fail because the model of health care work inscribed in these tools clashed too much with the actual nature of health care work* [2].

There is a clear need for a more timely and flexible development, validation, and deployment of CDS tools. In this paper, we will present an approach to allow for a more agile development of CDS tools with quicker cycles of development and user testing. The approach is based on:

- Ontology-based data integration to alleviate the problem of data heterogeneities which arise by the use of different standards, codifications and data schemas. The goal is to make more data sources more easily available for both development and validation.
- The use of data mining technologies to speed up the process of generating more and better medical knowledge. This shall provide both the developers of the systems and the medical experts with a data-driven approach to develop useful and understandable CDS models.
- The development of a flexible architecture of a CDS tool, in order to facilitate a faster validation and deployment of CDS tools in new hospitals.

We believe that such an approach may considerably speed up the implementation and validation cycles, and thus lead to a quicker development of more useful CDS tools.

The remainder of the paper is structured as follows: In Section II, we introduce the scenario on which the work in this paper is built. Section III presents the approach of ontology-based data annotation. Section IV discusses in more detail the challenges of clinical decision support, and based upon this introduces our architecture for a CDS tool. Section V describes how decision support models can be automatically learned from data. Section VI concludes.

II. BACKGROUND AND SCENARIO

A. Background

This work was carried out in the *p-medicine* project [3]. The goal of *p-medicine* is to take a step towards personalized medicine by building an infrastructure that addresses all tasks in executing clinical trials, from data gathering in hospitals over semantic integration, data warehousing, and data mining

This research has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under the project *p-medicine* (N^o270089).

¹Stefan Rüping is with Fraunhofer IAIS, 53754 St. Augustin, Germany stefan.rueping@iais.fraunhofer.de

²Alberto Anguita is with Biomedical Informatics Group, Universidad Politécnica de Madrid, Campus de Montegancedo, 28660 Madrid, Spain anguita@infomed.dia.fi.upm.es

³Anca Bucur is with Philips Research Eindhoven, The Netherlands anca.bucur@philips.com

⁴Traian Cristian Cirstea is with Philips Research Eindhoven, The Netherlands cristian.cirstea@philips.com

⁵Björn Jacobs is with Fraunhofer IAIS, 53754 St. Augustin, Germany bjoern.jacobs@iais.fraunhofer.de

⁶Antje Torge is with Department of Pediatrics, University Hospital Schleswig-Holstein, 24105 Kiel, Germany antje.torge@uksh.de

to produce new clinical knowledge that finally can be brought to the bedside, for example in a CDS system.

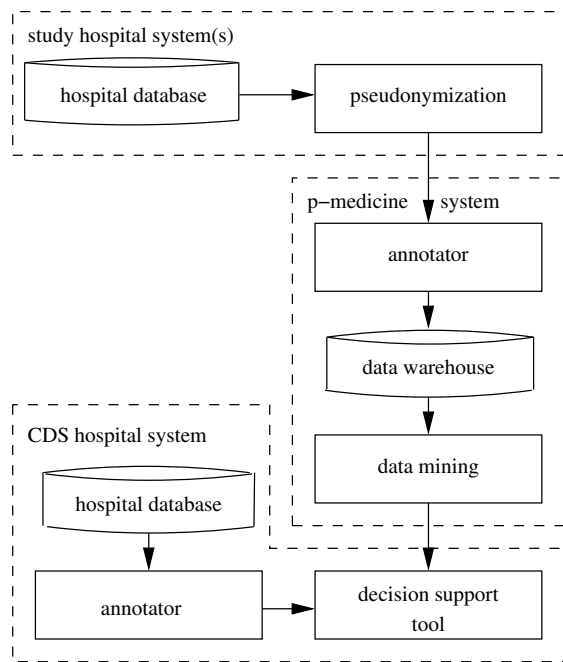


Fig. 1. Architecture diagram containing one or more hospital systems that provide data, the *p-medicine* system to process the data, and a CDS system in another hospital.

Figure 1 depicts the setup of our scenario. We assume that one or more hospitals decide to join in on a study to develop a new CDS model. The process runs as follows:

- 1) Data export: In each hospital, data from its hospital information system is exported in pseudonymized form.
- 2) Data processing: The data is imported into the centralized *p-medicine* system, where its fields are annotated with an ontology to facilitate their integration into a common schema in the data warehouse. Data mining tools analyze data to generate a CDS model.
- 3) Decision support: inside the systems of a new hospital, the learned decision support model will be applied to patient data. However, since the model contains rules which are expressed in terms of the selected ontology, another annotation step is necessary.

We believe that the implementation of this scenario can improve the creation of decision support models, because it makes better use of practical clinical experience in two ways, both from the simplified integration of new data sources from the study hospitals, and the quicker deployment in additional hospitals, which leads to better user feedback and validation.

B. Evaluation of CDS Models

The evaluation of CDS models is a complex task. Models are peer-reviewed by experts based on published results before being applied in practice. However, the more complex the model is, the harder such an evaluation is. We suggest to complement the manual evaluation with a statistical evaluation on new data sources. Being able to easily transfer the

model to new data sets from new hospitals is a key factor in faster testing of models on more data. In the end, this will facilitate to peer-review a model also on the data level.

C. The ALL-BFM Leukemia Study

Before we go into the technical details of our approach, let us shortly introduce a study on the subject of leukemia treatment, which will serve as our running example.

The most common malignant disease in childhood is the acute lymphoblastic leukemia (ALL). For effective therapy based on strictly controlled clinical data, the children are treated according to a standardized study protocol within Europe's largest therapy trial for the treatment of childhood ALL, the Berlin-Frankfurt-Münster study (ALL-BFM). The protocol includes three basic treatment arms according to patients with standard risk (SR), medium risk (MR) and high risk (HR) of relapse. For the most effective treatment with less side effects, it is essential to know the risk group as early as possible. The value of minimal residual disease (MRD) at treatment day 78 has been found out to be a strong predictive value for the outcome [4]. To calculate the risk of a patient earlier than that, the goal is to construct a predictive model for high MRD values – and therefore high risk of relapse – based only variables which are available very early.

From the data collected in the ALL-BFM trial a clinical data set of 60 variables was compiled. The data includes information on blood count at several time points, general genetic information, special information on the disease, physical characteristics and general condition, treatment and outcome. These are available for about 5000 patients.

III. DATA ANNOTATION

The Annotator is a tool in charge of homogenizing heterogeneous data sets from different hospitals. The tool is set in the context of a centralized data integration approach [5] which aims to offer biomedical researchers integrated access to multiple data sources. Availability of a great amount of data sources leads to more fruitful research. However, the ability of researchers of taking advantage of the integrated approach is greatly affected by heterogeneities among the data sources. The use of different standards, codifications and data schemas seriously limits the possibilities of accessing multiple repositories. The Annotator undertakes the process of homogenizing different schemas by exploring the semantic correspondences between them. The output of this tool is an annotation of the database which enables the automatic homogenization of the data into terms of a common vocabulary.

The *p-medicine* data integration layer relies on a Data Warehouse in charge of collecting the homogenized version of a set of heterogeneous data sources gathered from different hospitals and research institutions. Semantic interoperability is achieved through the use of HDOT, an ontology which describes the domain of health and biomedical research [6]. Hence, the Annotator tool produces semantic correspondences between the schemas of the hospital database systems and HDOT (it must be noted that the Annotator is not dependant on HDOT, and thus the approach presented in this

```

<physical_view><paths>
  <path>
    <class>#Patient</class>
    <property>#has</property>
    <class externalBound="Bound1">#BLAST_15
  </class>
  </path>
</paths></physical_view>
<conceptual_view><paths>
  <path>
    <class>#Biopsy</class>
    <property>#reveals</property>
    <class>#Cell</class>
    <class>#hasPercentage</class>
    <class externalBound="Bound1">#Percentage
  </class>
  </path>
  <path>
    <class>#Biopsy</class>
    <property>#afterTreatment</property>
    <class condition="#hasValue=15">#Delay
  </class>
  </path>
</paths></conceptual_view>

```

Fig. 2. XML snippet describing the BLAST.15 variable in HDOT.

paper is extensible to any other ontology). This is done by converting the data sources to RDF, graphically representing these data together with the ontology, and offering end-users an intuitive interface specifying the equivalences of elements.

In this paper, the Annotator was used to define semantic correspondences between the study variables and equivalent concepts in the ontology. Each study variable was mapped to equivalent variables in the ontology. For example, the field *BLAST_15* (percentage of blasts in bone marrow after 15 days of treatment) required navigating and constructing a view of the ontology which represented an equivalent concept. This view included the *patient* and *biopsy* concepts involved in the process of measuring the blast count in patients. Figure 2 shows a piece of the XML annotation file generated.

The resulting XML file contains the semantic relations of the two schemas. This annotation was used to translate the study data into the *p-medicine* data format. A second annotation was created for the CDS hospital data system and enables transfer of knowledge generated in the context of the *p-medicine* platform into this hospitals' data infrastructure.

IV. CLINICAL DECISION SUPPORT SYSTEMS

Knowledge and information maintenance plays an important part in the effectiveness and success of CDS systems. In addition to the maintenance of external knowledge and access to the latest evidence, the ability to efficiently update the CDS tools to take this new evidence into account is of utmost importance. A flexible and extendable CDS solution that is easy to update with the inclusion of new clinical evidence in the form of clinical models addressing specific decisions can address these challenges.

Our solution facilitates the introduction of new clinical models, access through standard interfaces and supports easy

updates and extensions that can deal with the high rate of change in medical knowledge and can also lead to higher acceptance of both the CDS tool and the underlying clinical models by the clinical users. Another advantage is the ability to independently validate the clinical models, gradually add them to the CDS tool, and apply them in patient care.

Workflow integration is another key success factor for the CDS systems. To facilitate adoption of the proposed models, we make sure that the CDS tool encapsulating the clinical models fits in the current clinical workflow and can be efficiently used in the care setting. This ensures that all extensions and updates with new models will be easily integrated in the current way of working of the clinicians.

The *p-medicine* CDS environment aims to enable clinicians to receive advice based on the knowledge models discovered through data mining of the available clinical trial datasets. The decision support system, from an architectural point of view, consists of: (1) a user interface, (2) a database to store the clinical models, and (3) a set of one or more workers that actually run the models and generate advices. This architecture makes the system easily scalable. If more models are required, more engines can be added easily.

The decision support system allows scientists to implement their clinical models in various programming languages. Currently, the Jess engine¹ is integrated into the platform, as it is powerful and flexible enough to cope with most of the requirements of clinical models. Jess is a rule based engine that uses an enhanced version of the Rete algorithm [7] to process rules. Basically, Jess is an application that continuously applies a set of if-then statements (rules) to a dataset, and outputs the results to the caller.

As data is stored by different hospitals by making use of different standards (or proprietary solutions), modifications and data schemas, the access to data is not standardized. Therefore, feeding the right data to the model is not a trivial process. The decision support system solves this apparently difficult situation by separating the clinical model from the actual location of the data. It acts as an abstraction layer between the two, and the actual binding between the input variable of the model and the actual database fields of the hospital is performed automatically by making use of the Ontology Annotator mapping XML file.

V. DATA MINING TO CONSTRUCT CDS MODELS

Data Mining has long since been used successfully in the construction of clinical decision support systems [8]. Of special interests for medical applications are knowledge discovery approaches which can learn understandable rules from data, in order to ease the process of generating guidelines which are understandable to the physician [9].

Subgroup discovery [10] is a well-known approach for finding patterns in data that are not only predictive, but also understandable to the user. The basic idea of subgroup discovery is to investigate all subsets of a database which can be described by an if-then-rule, calculate a measure of

¹<http://www.jessrules.com>

```

(defrule RULE_0
  (MAIN::input_variable BLAST_15 ?BLAST_15&:(>= ?BLAST_15 58.5)&:(<= ?BLAST_15 96.0))
  (MAIN::input_variable IMMUNOM ?IMMUNOM&:(= ?IMMUNOM 3))
  (MAIN::input_variable TEL_AML ?TEL_AML&:(= ?TEL_AML 1))
  =>
  (assert (MAIN::output_variable RULE_0_MATCH TRUE)))

```

Fig. 3. Rule definition in the form of Jess script.

statistical deviation of a variable of interest in this subset from the overall data set, and find those subsets with the statistical most significant deviation. The advantage of using if-then-rules to define subgroups is that users often find it easy to inspect and understand these formulas, as they can be easily cast to natural language sentences.

What distinguishes subgroup discovery from most other machine learning approaches is that it does not require to find a model that explains the complete database, but only those parts of the data that can be explained with high certainty (i.e. high statistical significance). This property makes them well suited for tasks of decision support, because here it is required that the recommendation that is given to the treating physician is correct with high certainty, but it is not required that in all cases a recommendation must be given.

To give an example, the following rule was found on the ALL-BFM data set: $BLAST_{15} \in [58.5, 96] \ \& \ IMMUNOM=3 \ \& \ TEL_{AML}=1 \Rightarrow P(HR) = 60.8\%$. It is based on three variables, $BLAST_{15}$ (amount of blasts in the bone marrow at treatment day 15 in %), $IMMUNOM$ (patient has expressed myeloid antigens) and TEL_{AML} ($TEL/AML1$ negative patient). It predicts that with a certain combination of these variables the probability of being a high risk patient rises from its original 7.1% to 60.8%. As the rule just serves as a technical example, the medical importance is of minor interest in this context. However, we would like to point out that this is in line with previous research that has established the relevance of these factors individually [11]–[13]

The interesting property of rules in this form is that it is easily possible to compile these rules into programs of a given programming language, such as Jess, for an automated execution of the models. To this end, a parser application produces an logically equivalent Jess script file from a subgroup model. The parser performs an pre-generation analysis of the given rules to prevent code-duplication, identify shared input variables and collect information about the number of rules to be parsed. The following steps are executed by the parser:

- For each rule one output variable needs to be declared. These output variables will be transferred to the surrounding environment and indicate whether a rule fired.
- For each rule one Jess rule definition is generated.
- The rules' left-hand-side is populated with Boolean conditions from the given rules. This is achieved by binding the corresponding input data to a local variable on which then functional checks are performed. If all left-hand-side conditions evaluate to true then the right-hand-side asserts the value TRUE to the rule's output.
- To create a valid Jess script, surrounding statements

need to be inserted that declare the context and execute the rule engine when the script is started. Also, metadata about the input and output variables is generated.

The Jess script of our exemplary rule is given in Figure 3.

VI. CONCLUSIONS

Developing tools for clinical decision support remains to be a challenging problem. We have presented an approach to make the development and validation process more timely and flexible. The approach is based on three pillars, namely (1) making the integration of new clinical databases, – hence the empirical basis of model development and validation – easier by the use of ontology-based annotation, (2) the use of appropriate data mining approaches to simplify the data-driven construction of CDS models in a standardized language, and (3) the implementation of a flexible CDS environment based on standardized data and model descriptions.

REFERENCES

- [1] Morris, "Algorithm-based decision making," in *Principles and Practices of Intensive Care Monitoring*, Tobin, Ed. New York: McGraw-Hill, 1998, pp. 1355–1381.
- [2] Wears R.L., "Computer technology and clinical work: still waiting for godot," *The Journal of the American Medical Association*, p. 12611263, 2010.
- [3] Graf, Anguita, Bucur, Burke, Claerhout, Coveney, *et al.*, "p-medicine from data sharing and integration via vph tools to personalized medicine," in *Proc. VPH2012*, London, 2012.
- [4] Dongen, Seriu, Panzer-Grümayer, Biondi, Pongers-Willemsse, Corral, *et al.*, "Prognostic value of minimal residual disease in acute lymphoblastic leukaemia in childhood," *Lancet*, vol. 352, pp. 1731–1738, 1998.
- [5] Sujansky, "Heterogeneous database integration in biomedicine," *Journal of Biomedical Informatics*, vol. 34, pp. 285–298, 2001.
- [6] Sanfilippo, Schwarz, and Schneider, "The health data ontology trunk (HDOT): towards an ontological representation of cancer-related knowledge," in *Proc. IARWISOCI2012*, Athens, 2012.
- [7] Forgy, "Rete: A fast algorithm for the many pattern/many object pattern match problem," *Art. Intelligence*, vol. 19, pp. 17–37, 1982.
- [8] Soo, Wang, and Wang, "Learning and discovery from a clinical database: an incremental concept formation approach," *Artificial Intelligence in Medicine*, vol. 6, pp. 249–261, 1994.
- [9] Morik, Imhoff, Brockhausen, Joachims, and Gather, "Knowledge discovery and knowledge validation in intensive care," *Artificial Intelligence in Medicine*, vol. 19, no. 3, pp. 225–249, 2000.
- [10] Klösgen, "Explora: A multipattern and multistrategy discovery assistant," in *Advances in Knowledge Discovery and Data Mining*. MIT Press, 1996, pp. 249–271.
- [11] Schrappe, Reiter, Zimmermann, Harbott, Ludwig, Henze, *et al.*, "Longterm results of four consecutive trials in childhood all performed by the all-bfm study group from 1981 to 1995," *Leukemia*, vol. 14, pp. 2205–2222, 2000.
- [12] Wiersma, Ortega, Sobel, and Weinberg, "Clinical importance of myeloid-antigen expression in acute lymphoblastic leukemia of childhood," *Engl J Med*, vol. 324, pp. 800–808, 1991.
- [13] Loh, Silverman, Young, Neuberger, Golub, Sallan, *et al.*, "Incidence of tel/amll1 fusion in children with relapsed acute lymphoblastic leukemia," *Blood*, vol. 15, pp. 4792–4797, 1998.