Digital Healthcare Empowering Europeans
R. Cornet et al. (Eds.)

doi:10.3233/978-1-61499-512-8-185

An IT Architecture for Systems Medicine

Matthias GANZINGER[1], Matthias GIETZELT, Christian KARMEN, Daniel FIRNKORN, and Petra KNAUP
University of Heidelberg, Institute of Medical Biometry and Informatics, Heidelberg, Germany

Abstract. Systems medicine aims to support treatment of complex diseases like cancer by integrating all available data for the disease. To provide such a decision support in clinical practice, a suitable IT architecture is necessary. We suggest a generic architecture comprised of the following three layers: data representation, decision support, and user interface. For the systems medicine research project "Clinically-applicable, omics-based assessment of survival, side effects, and targets in multiple myeloma" (CLIOMMICS) we developed a concrete instance of the generic architecture. We use i2b2 for representing the harmonized data. Since no deterministic model exists for multiple myeloma we use case-based reasoning for decision support. For clinical practice, visualizations of the results must be intuitive and clear. At the same time, they must communicate the uncertainty immanent in stochastic processes. Thus, we develop a specific user interface for systems medicine based on the web portal software Liferay.

Keywords. Clinical Decision Support Systems, Individualized Medicine, Statistical Models

Introduction

Systems medicine is a current approach to find new ways of targeting complex diseases like cancers. The concept of systems medicine is often defined as the application of systems biology strategies to medicine [1, 2]. However, systems medicine is considered not only a way of describing functional aspects of the disease but also the support of a treatment decision for the patient. Thus, support by specific information technology (IT) systems is necessary. To provide a systems view to a specific disease it is essential to integrate all available data. Often, these data come from different sources with different, not standardized data formats. Certainly, omics data will play a substantial role in systems medicine data collections. They are complemented by phenotype data, such as clinical data or data from clinical trials. Since these data come from different sources, they might describe the same entity differently in terms of data structure and semantics. To make data available for a comprehensive view on the disease, harmonization is necessary.

Based on such a unified data collection, models specific for the disease are built. For many diseases the construction of deterministic models is unlikely. Thus, stochastic models containing a substantial amount of uncertainty will predominate. Essentially, an

[1] Corresponding Author: Matthias Ganzinger, University of Heidelberg, Institute of Medical Biometry and Informatics, Im Neuenheimer Feld 305, 69120 Heidelberg, Germany. Email: matthias.ganzinger@med.uni-heidelberg.de

IT system is necessary to support this systems medicine process as a whole. On the one hand, the system has to support researchers investigating the disease. This can be done for example by providing access to the complete data collection by means of a research data warehouse. Based on these data, researchers can develop new models for improving decision support. On the other hand, clinical decision making should be supported appropriately. Based on parameters for a patient, an individual treatment decision should be supported in accordance with the patient's treatment goals. With respect to timely constraints in daily routine, complex dependencies should be displayed in a clear and intuitive way.

In this manuscript, we propose a generic IT architecture for systems medicine. A concrete instance of the architecture is implemented for the systems medicine research project "Clinically-applicable, omics-based assessment of survival, side effects, and targets in multiple myeloma" (CLIOMMICS). Multiple Myeloma (MM) is a cancer of plasma cells producing monoclonal antibodies and accumulating in bone marrow. Incidence of MM is 4 to 6 per 100.000 people per year with a median age at diagnosis of 65 to 70 years [3]. Currently, only a subset of the myeloma patients responds sufficiently to the therapy provided. In addition, some patients suffer from severe side effects. However, no method of predicting response or side effects is available so far. The aim of the project is to integrate omics-data with conventional clinical and molecular data into clinical routine to provide a decisions system for the best therapy for individual patients. The decision support system should help physicians maximizing efficacy of therapy while minimizing side effects. The architecture implemented for CLIOMMICS can be used in other disease contexts as well.

1. Methods

For systems medicine applications, we propose a three-layer-architecture. Since systems medicine is based on the integration of all available data, the first layer covers the representation of data collected and harmonized for systems medicine purposes. Technically, this can be for example a database, a research data warehouse (RDW), or a federated data pool.

The second layer covers decision support for systems medicine. Here, data from the first layer are processed according to the disease specific model. The model might be derived from data of the first layer or it might be based on external knowledge like an ontology. A combination of both is possible as well.

The third layer is the user interface. Clinicians and researchers have different requirements when using the system and thus need different interfaces. Researchers need access to layer one data since they generate the knowledge necessary for parameterizing decision support in layer two. Clinicians on the other hand need an interface providing access to the results of the decision support system in a comprehensible way.

2. Results

In this section we describe the concrete systems medicine IT architecture we developed for CLIOMMICS. An overview of the architecture is shown in Figure 1. This figure also provides a mapping between modules of the CLIOMMICS architecture and the

generic three layer model. In the following paragraphs we discuss the modules of the architecture in detail.

2.1. Layer 1: Data Representation

Research Data Warehouse: One of the core components in our architecture is the RDW based on i2b2 [4]. It hosts all phenotype specific data available for the definition of the disease. For our project on MM, these data are comprised of clinical routine data and data from clinical trials that were collected for two decades. Since they originate from different sources, data harmonization was inevitable. For harmonization, we applied our data harmonization framework [5]. This framework includes technical support for harmonization including a mapping tool and generic extract, transform and load (ETL) processes for loading the harmonized data into the RDW. The RDW directly provides a user interface for researchers to access data for research purposes. From the patient cohort, sub-cohorts with specific properties are filtered and exported. Detailed statistical analysis is performed with dedicated tools like R or SAS.

Generic Case Extractor (GCE): Data in the RDW are organized according to the star schema. While this data schema is optimized for analytical queries, for some purposes a flat case-oriented presentation of the data is desirable. We implement a generic component for i2b2 allowing a comprehensive data export as a matrix containing one line per case. This component both supports the data export via the research interface and to downstream component *Parameter Selection Engine (PSE)*.

Omics Data Store and Omics Pipeline: Since omics data tend to have a large volume, they are not suited for inclusion into the RDW. Thus, our architecture includes file storage for omics data like gene expression or single nucleotide polymorphism data. These data are handled in specific omics pipelines. Only results like the state of disease associated gene sets are added to the RDW to complement patients' clinical data.

2.2. Layer 2: Decision Support

Parameter Selection Engine: PSE performs analyses on the case-oriented data sets

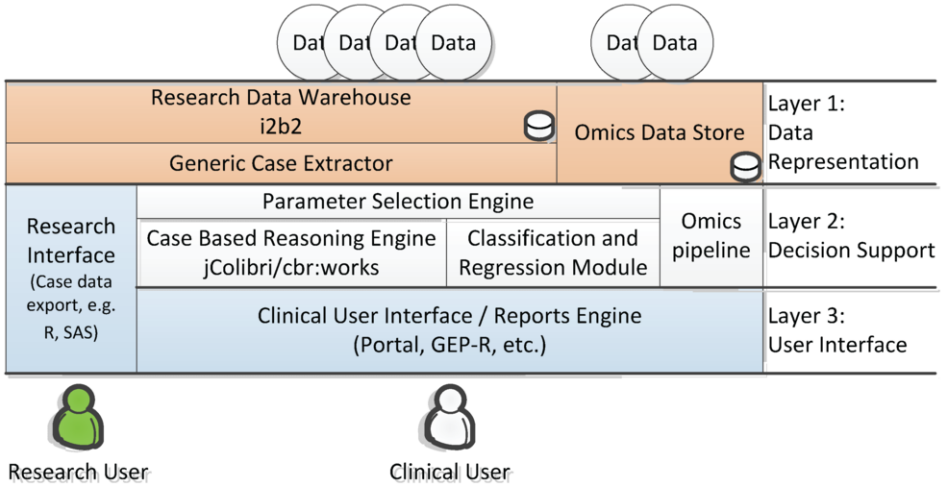

Figure 1. Architecture of the CLIOMMICS systems medicine application

provided by GCE. It is responsible for selecting a subset of parameters explaining a desired endpoint best. Here, the Weka implementation of the CfsSubsetEval-algorithm is used [6]. The parameter set is used for adjusting weight parameters and distance measures for case-based reasoning. Further, it provides input to the C*lassification and Regression Module (CRM)*.

Case-Based Reasoning Engine: For the support of treatment decisions we follow a case-based reasoning (CBR) approach [7]. The CBR module hosts the case base derived from RDW and omics data. Thus, it provides a historic view on the patients treated for the disease in question, in our case MM. Together with the similarity and weighting measures, the case base provides a stochastic model of the disease. Based on the clinical and omics parameters of a newly admitted patient, similar cases are selected from the case base. The relevant outcome depends on the treatment goals for patient: While commonly best overall survival is desired, minimizing side effects and thus optimizing quality of life might be an alternative strategy.

Classification and Regression Module: The CRM automatically builds prediction models regarding an endpoint of interest. Endpoints of interest are expected response to treatment, prediction of overall and progression-free survival, and likelihood of side effects. Thereby, the CRM makes use of the random forest implementation in Weka [8]. The selected parameters of the PSE were either classified, if the chosen endpoint is nominal or it computes a regression model, if the endpoint is numeric. Then, the model will be evaluated using a separate test set and additionally regarding the training set to measure the goodness-of-fit.

2.3. Layer 3: User Interface

The establishment and user acceptance of a systems medicine application strongly depends on an optimized user interface. Complex dependencies should be presented in a way allowing clinical consequence to be drawn within a suitable time frame. We use the web-based portal system *Liferay* [9] for the user interface layer. The user interface interacts with the CBR engine and the CRM to acquire treatment alternatives and likelihoods in case the desired outcome can be achieved with different lines of treatment. In addition, the user interface provides a workflow component for generating reports for communication among health care providers [10].

3. Discussion

We provide an IT architecture to be used for systems medicine. In our research project CLIOMMICS we successfully implemented the RDW component by using i2b2. By now it contains roughly 300 parameters for around 1000 patients. Corresponding ETL processes are established to load the RDW with clinical data. Further, the omics data store with 904 gene expression microarray data sets and corresponding pipeline are available for the project. To utilize the data for research and decision making we implemented the GCE tool for a case oriented view. Currently we are in the process of establishing the CBR, CRM, and clinical user interface components. The system integrates several existing open-source software products. Likewise, we plan to publish the modules developed by ourselves under an open source license for free.

The core element of our architecture in terms of prognosis and decision support is the CBR module. We expect this approach to provide a systems perspective in decision

support, since all available data can be integrated into the decision making process. However, it is very important to choose similarity and weighting measures with great care. For example, attributes of the cases will partially come from omics data sets. We expect a significant amount of noise induced by this high dimensional data source that needs to be filtered out. This problem is caused by the nature of the data involved.

An alternative platform designed for systems medicine might be G-DOC [11]. Like our architecture, it covers both clinical and omics data. However, the system is a tool to perform powerful analyses for research and to some extend in clinical practice with manual workflows.

The generic architecture for systems medicine applications we describe in this document provides building blocks that can be applied to other systems medicine projects in the community. Due to its modular design, components can be replaced if necessary. For example, the CBR module might be changed if a more precise model for a specific disease exists. In the future, the community might agree on standardized interfaces between the layers of the architecture which might lead to a resource pool of components for systems medicine. As a next step after completing the implementation of our architecture, we plan to validate it for MM. Further, we plan to apply it to other cancerous diseases like lung cancer to proof the general applicability of the architecture.

Acknowledgement

CLIOMMICS is funded by the German Ministry of Education and Research within the e:Med initiative. Grant id: 01ZX1309A

References

[1] Kaminsky DA, Irvin CG, Sterk PJ. Complex systems in pulmonary medicine: a systems biology approach to lung disease. J Appl Physiol 2011; 110(6):1716–22.
[2] Tian Q, Price ND, Hood L. Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine. J Intern Med 2012; 271(2):111–21.
[3] Harousseau J, Moreau P. Autologous hematopoietic stem-cell transplantation for multiple myeloma. N Engl J Med 2009; 360(25):2645–54.
[4] Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc 2007:548–52.
[5] Karmen C, Ganzinger M, Kohl CD, Firnkorn D, Knaup-Gregori P. A framework for integrating heterogeneous clinical data for a disease area into a central data warehouse. Stud Health Technol Inform 2014; 205:1060–4.
[6] Hall MA. Correlation-based Feature Selection for Machine Learning. Hamilton, New Zealand: University of Waikato; 1998.
[7] Aamodt A, Plaza E. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications 1994; 7(1):39–59.
[8] Breiman L. Random Forests. Machine Learning 2001; 45(1):5–32.
[9] Sezov R, Jr. Liferay in action: The official guide to Liferay portal development. Shelter Island: Manning; 2012. (Official guide).
[10] Meissner T, Seckinger A, Rème T, Hielscher T, Möhler T, Neben K et al. Gene expression profiling in multiple myeloma--reporting of entities, risk, and targets in clinical routine. Clin Cancer Res 2011; 17(23):7240–7.
[11] Madhavan S, Gusev Y, Harris M, Tanenbaum DM, Gauba R, Bhuvaneshwar K et al. G-DOC: A Systems Medicine Platform for Personalized Oncology. Neoplasia 2011; 13(9):771–83.