Health Information Research Platform Heidelberg

In-Memory-based and Entity-Attribute-Value-based Real-Time Data Analyses as part of a Clinical Research Infrastructure

Gerd SCHNEIDER[a,1], Harald AAMOT[b], Björn SCHREIWEIS[a], Nina BOUGATF[c] and Björn BERGH[a]
[a] *Center for Information Technology and Medical Engineering, Heidelberg University Hospital, Germany*
[b] *NCT Trial Center, German Cancer Research Center, Heidelberg, Germany*
[c] *Department of Radiation Oncology, Heidelberg University Hospital, Germany*

Keywords. Data Warehouse, Entity-Attribute-Value Data Model, Real-Time Data Analyses, In-Memory, Data Protection

Introduction

Next-generation-sequencing technologies and various other methods gain data from human biospecimen both for translational research and personalized medicine. The combination of gene expression data with clinical data and the interdisciplinary collaboration of physicians, molecular biologists and computer scientists could be the key to new innovations in translational research. The integration of highly structured genetic data with predominantly unstructured clinical data imposes a big challenge, especially as unstructured clinical data can be very heterogeneous and distributed over multiple source systems. Thus, a structured clinical data warehouse is a prerequisite to be able to combine genotypes with phenotypes for care-oriented as well as research-oriented data analyses, while complying with strict data protection guidelines.

1. Methods

Driven by an oncological use case the Heidelberg University Hospital and the National Center for Tumor Diseases (NCT) developed the vision of a Health Information Research Platform (HIReP) serving as a single point-of-truth for storage and analyses of clinical data for care and research. Besides a central data warehouse, the HIReP vision comprises various support components like, for example, identity and consent management, and anonymization and pseudonymization services in case of research-oriented data usage. As an initial step, we decided to set up the central clinical data warehouse by extracting clinical data from various clinical information systems (hospital information system, cancer registry and an application-specific database for structured documentation of radiation therapies). Subsequently, the extracted data are transformed into an entity-attribute-value data model and loaded into a central column-oriented clinical data warehouse. As part of the oncological use case not only structured clinical data, but also doctors' letters are extracted and analyzed for biomarker information with text analysis and semantic knowledge representation and retrieval methods. The biomarker information extracted from the letters is also loaded into the central clinical data warehouse.

2. Results

In a prototype we were able to show that clinical information can be loaded into the central data warehouse and be visualized with a graphical, HTML 5-based frontend. Responses to user formulated queries show up in real-time due to the column-oriented in-memory database technology used in the project. Additionally, we were able to show that information extraction from unstructured documents is possible. A manual validation by an oncological physician showed a high quality in the information extraction results. Also negations and other semantically challenging concepts were handled by the text analysis.

3. Discussion and Outlook

The realization of our HIReP vision and concept is still ongoing. Especially the support components for identity management, consent management and anonymization/pseudonymization have to be further established as these components are mandatory with regard to data privacy-compliant usage of clinical data in a research context. As soon as these components are available we will analyze and promote the interlinkage of the structured data warehouse content with genetic data in a privacy preserving way. The results will be validated together with the respective clinicians. In principle, our decision to use a column-oriented, in-memory database as technological basis for the clinical data warehouse is still unchallenged. Within the current oncological setting it can handle millions of entries, e.g. diagnoses and laboratory values, without a perceivable delay. The main bottleneck by now is the extract-transform-load process from the information sources to the data warehouse. For this reason a delta loading mechanism is planned and the advantages of Lambda architecture for Big Data systems are investigated. Besides this, a small group of researchers analyzes the commonalities and differences between the HIReP approach and the i2b2 approach.

[1] Corresponding Author: Speyerer Straße 4, 69115 Heidelberg, Germany. gerd.schneider@med.uni-heidelberg.de