MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220237

Towards Data Analysis Environment for Multi-Partner Clinical Research Project - SMART-CARE

Aleksei Dudchenko, Friedemann Ringwald, Petra Knaup, and Matthias Ganzinger

Institute of Medical Informatics, Heidelberg University Hospital, Heidelberg, Germany

Abstract

A Systems Medicine Approach to Stratification of Cancer Recurrence (SMART-CARE) establishes mass spectrometry-based systems medicine technologies and data analysis pipelines employing expertise of the multiple partners from Heidelberg biomedical campus. We have established a central linked data repository that links clinical, mass spectrometry, and data analysis teams to enable a full cycle of data management. Other questions of setting up the data analysis environment for the multi-partner clinical research project are addressed in this work, too.

Keywords:
Data analysis, system medicine, cancer recurrence.

Introduction

Cancer recurrence remains the main reason for cancer-related death. Stratification of cancer patients and personalized cancer treatment that relies on genomic data has improved cancer patient stratification and the development of novel therapies. Nevertheless, other important layers, such as the proteome and metabolome could be combined with rapid progress in the development of artificial intelligence and machine learning and may improve personalized medicine and bring it to a new level. This requires a systems medicine approach that can only be achieved by tight integration of clinical, technological, and computational expertise and resources. SMART-CARE brings together clinical, mass spectrometry and computational expertise on the Heidelberg biomedical campus to firmly establish mass spectrometry-based systems medicine technologies and data analysis pipelines.

A key aspect of SMART-CARE is to link proteomic and metabolomic data with available genomic and clinical information for integrative data analysis. This requires standardization and semantic harmonization of the metadata of all sources and a central repository to connect multidimensional and heterogeneous data types.

This is technically and logistically highly challenging, since it requires various criteria that need to be met simultaneously. These criteria include a scientific environment that facilitates coordinated integration; standardized procedures for sample preparation and mass spectrometry for robust proteome and metabolome analysis; and appropriate means for clinical validation and implementation (e.g. trials, multidisciplinary teams for personalized medicine).

From the data analysis perspective the criteria are 1) availability of a significant number of well-annotated clinical samples that adhere to defined clinical inclusion criteria and that are collected by standardized procedures; 2) established advanced IT infrastructure, providing run-time environment, integration of third-party analytical tools and tools for complex data integration.

This article is devoted to aspects of development of data analysis environment for the multi-partner clinical research project - SMART-CARE.

Methods

To ensure the fulfillment of the assigned tasks of the project we have built the central SMART-CARE linked data repository (LDR) that will serve as a central and uniform link between clinical, mass spectrometry, and data analysis teams for extensive data mining.

Collaborative partners and roles

Three types of partners participate in the SMART-CARE LDR based on OpenBIS software [1]: clinical partners, proteomics and metabolomics mass spectrometry laboratories, and analytics groups. A data management team is responsible for administration and management of OpenBIS.

Clinical partners provide basic patients' data, sample data, biological and medical data. The data can be submitted to the central repository either by filling out data definition forms or via ETL (Extract, Transform, Load) processes.

As the next step, mass spectrometry partners submit laboratory data. Data definitions for those steps are also predefined depending on the sample type and type of the experiment. Again, data from labs can be submitted by filling out a form via the OpenBIS web interface or by uploading batch files.

The analytics groups can access acquired data for statistical analysis as well as for building, training, and running models. Results of the analysis are also stored in the central data repository and might be accessed by the corresponding clinic.

Data management challenges

Among the tasks to of the project, the following are especially important in data management prospective and require nontrivial solutions: 1) developing data definitions for clinical data. This should be in adherence to the HiGHmed [2] data definitions and archetypes; 2) developing data definitions for metadata collected from heterogeneous sources and different collaborative partners; 3) designing an identity management concept including a pseudonymization concept and SMART-CARE wide unambiguous IDs; 4) designing a concept for the data access, usage, and data protection that guarantees an easy and safe access to the harmonized and pseudonymized data sets for clinical, and data analysis teams.

Data analysis challenges

To enable multiple data analysis groups to perform analytical workflows we have identified the following tasks: 1) providing straight forward access to metadata, data, and raw data for all stored objects and types; 2) providing runtime environment for analytical software; 3) integration with third-party data analysis tools and solutions; 4) establishing environment for data-analysis pipelines management.

Results

Physical infrastructure

Since data to be stored in the central data repository include not only pre-processed data but also raw data, which reaches several gigabytes per file, a dedicated high-performance server was established. The SMART-CARE server ensures a sustainable and secure access to the central data platform. To provide runtime environment for training ML models, the SMART-Care server is equipped with powerful GPUs.

Data definitions

At the current stage, some data definitions (OpenBIS masks) are still pending since changing them after being deployed at the live system is not recommended. Nevertheless, we have fully completed set of data definitions for Hematology Space of SMART-CARE OpenBIS instance (operated by Department of Internal Medicine V, Heidelberg University Hospital). The full cycle of data collection for this clinical partner has been implemented in the live system. First real-live samples and aliquots data have been gathered into the system.

SMART-Care currently involves four different clinical partners. Data submitted by the clinics are not homogeneous, a different data definition should be applied in this case. To simplify submitting and accessing the data at OpenBIS, we established distinct spaces for the clinical partners. Spaces limit and define the availability of data and OpenBIS objects.

Data analysis requirements

Having starting collecting real data and finalizing missing data definitions, we run the active phase of clarifying and specifying requirements of data analysis groups.

In very broad overview, data analysis require three components: data and access to it, models and data analytical pipelines with possibility to modify them, and an environment to run the pipelines with various subsets of the data. Each of the components opens questions and needs decisions made jointly with all analytical groups of the SMART-CARE consortium. The decisions must also take into account interests of clinical and mass-spectrometry partners.

Discussion and Conclusion

The question of a clinical data definition is not finalized yet. Despite the fact that we have defined a set of data to be provided, this set might not be optimal for the purposes of further data analysis. Data definitions, or masks in OpenBIS terminology, have been obtained in tight cooperation with all partners. However, further changes and refinements may be required, and these are not always easy to perform.

Some parts of the first-level raw data produced by the laboratory equipment can only be interpreted by specialized software from the vendor of the equipment. Does storing such data provide a meaningful benefit that is worth occupying several gigabytes of space per file?

LDR provides access to large amounts of data that has great scientific potential. We plan to make it possible to build, train, and run predictive models and analytical pipelines onside, without downloading the data This becomes more relevant if we consider the possibility for researchers outside of the SMART-CARE consortium to use the data for analytical purposes, but not to let them have direct and full access to the data.

Data management and analysis solutions reported in this article bring together clinical, mass spectrometry, and data analysis partners enabling the improvement of personalized medicine in cancer treatment.

References

[1] A. Bauch, I. Adamczyk, P. Buczek, F.J. Elmer, K. Enimanev, P. Glyzewski, M. Kohler, T. Pylak, A. Quandt, C. Ramakrishnan, C. Beisel, L. Malmström, R. Aebersold, and B. Rinn, OpenBIS: A flexible framework for managing and analyzing complex data in biology research, *BMC Bioinformatics*. **12** (2011) 1–19. doi:10.1186/1471-2105-12-468.

[2] B. Haarbrandt, B. Schreiweis, S. Rey, U. Sax, S. Scheithauer, O. Rienhoff, P. Knaup-Gregori, U. Bavendiek, C. Dieterich, B. Brors, I. Kraus, C.M. Thoms, D. Jäger, V. Ellenrieder, B. Bergh, R. Yahyapour, R. Eils, H.Gh. Consortium, and M. Marschollek, HiGHmed - An Open Platform Approach to Enhance Care and Research across Institutional Boundaries, *Methods Inf. Med.* **57** (2018) e66–e81. doi:10.3414/ME18-02-0002.

Address for correspondence

Corresponding author: Aleksei Dudchenko,
aleksei.dudchenko@med.uni-heidelberg.de