J. Mantas et al. (Eds.)

doi:10.3233/SHTI200563

Towards Intelligent Integration and Sharing of Stem Cell Research Data

Kirill BORZIAK [a], Tianye QI [a], John Erol EVANGELISTA [a], Daniel J.B. CLARKE [a], Avi MA'AYAN[a] and Joseph FINKELSTEIN [a,1]
[a]*Icahn School of Medicine at Mount Sinai, New York, NY 10029 USA*

Abstract. Advancements in regenerative medicine have brought to the fore the need for increased standardization and sharing of stem cell product characterization to help drive these innovative interventions toward public availability. Although numerous attempts have been made to store this data, there is still a lack of a platform that incorporates heterogeneous stem cell information into a harmonized project-based framework. The aim of this project was to introduce and pilot-test an intelligent informatics solution which integrates diverse stem cell product characteristics with study subject and omics information. In the resulting platform, heterogeneous data is validated using predefined ontologies and stored in a NoSQL repository. Pilot-testing was performed on nine sponsored stem cell projects conducting preclinical and intervention evaluations. The pilot-testing demonstrated the robustness of the proposed platform, by seamlessly harmonizing diverse common data elements, and the potential of this platform for driving knowledge generation from the aggregation of this shared data.

Keywords. Data aggregation, regenerative medicine, stem cells, common data elements

1. Introduction

Regenerative medicine is a highly innovative field that holds great promise for treating and even curing a variety of injuries and diseases by repairing, replacing, or regenerating damaged cells, tissues and organs [1]. To accelerate this field by supporting clinical research on adult stem cells while promoting the highest standards of scientific research and patient safety protection, the National Institutes of Health (NIH) has established, with the coordination of the Food and Drug Administration (FDA), the Regenerative Medicine Innovation Catalyst (RMIC). The main aim of this regenerative medicine research is to identify and promote stem cell therapies that lead to introduction of new treatments of previously incurable health conditions. To facilitate that goal, we aimed at developing a platform to store and disseminate diverse stem cell product information, clinical and pre-clinical trial outcomes, and study subject information in a harmonized and interchangeable way, with the aim of promoting novel knowledge generation through data sharing. The resulting platform should increase scientific rigor, accelerate progress, and stimulate innovation and partnership by creating an ecosystem that will promote collaboration and further

[1] Corresponding Author, Joseph FINKELSTEIN, Icahn School of Medicine at Mount Sinai, New York, NY 10029 USA; E-mail: Joseph.Finkelstein@mssm.edu.

improve the standards for characterizing stem cell products. The agile and flexible nature of such a platform, based on intelligent integrative informatics approaches, allows to seamlessly integrate diverse data, while maintaining stringent quality control measures. Use of data harmonization and common data elements (CDE) mapped to cross-linked biomedical ontologies allow the search and visualization of the heterogeneous data across multiple studies and to be quickly queried and yield results that can aid directly in improving the state of regenerative medicine. The novel contributions of this paper to the regenerative medicine and stem cell community are a scalable NoSQL Data Repository for storing all data generated as part of stem cell projects, programmatic interfaces (APIs) allowing for granular and secure access and depositing of all relevant information in the platform, and an intuitive user interface and search functions allowing for fine-tuned querying of all information in the platform. The aim of this article is (1) to describe the building blocks of our innovative data management platform, (2) to review result of the platform pilot-testing, and (3) to summarize and present an outlook for further work.

2. Methods

2.1. Platform Architecture

Our platform, called **Re**generative **Me**dicine **D**ata Repositor**y** (ReMeDy) is an implementation of the Signature Commons, a BD2K-LINCS [2] DCIC platform implemented through Docker and designed to store and search diverse metadata in an agile and flexible manner [3]. A Signature Commons instance was installed on a Linux server using the default installation instructions. The upload interface was built using ReactJS and Spring Boot.

2.2. Data Collection

Data for ReMeDy pilot-testing, in the format of the previously described multi-modular Common Data Elements (CDE) framework [4-5], was obtained from eight projects recently funded by the National Institutes of Health, plus a previously completed clinical trial, "Intramyocardial Injection of Mesenchymal Precursor Cells and Successful Temporary Weaning From Left Ventricular Assist Device Support in Patients With Advanced Heart Failure: A Randomized Clinical Trial", that was part of the Network for Cardiothoracic Surgical Investigations in Cardiovascular Medicine grant (Project Number: 7U01HL088942-03). The multimodular CDE template for stem cell research data sharing was populated using information available through NIH RePORTER and relevant papers published by the awarded project investigators. A representative set of 30 patients from trial 7U01HL088942-03 were included to populate the study subjects' section of ReMeDy. This included baseline, outcome, adverse events and readmission information previously submitted to BioLINCC.

3. Results

All data stored within a Signature Commons instance is designed to be organized hierarchically into three layers: resources, libraries, and signatures. For the ReMeDy implementation, resources represent the individual RMIC funded projects. Each project has one library, which contains the static CDEs, i.e., those that are not expected to change across the time course of the project, such as project information, cell product manufacturing and production information, and in-depth product characterization. Further information on the structure and organization of the CDEs contained in the libraries is provided below in the Modular CDE Framework section. Signatures represent individual study subject records and contain demographic, baseline, outcome and adverse event CDEs. As the study subject information will be updated on a periodic basis, partitioning it apart from the static CDEs will allow for easier and more reliable maintenance and retrieval of CDEs through the search functions.

3.1. Data Upload Interface

To improve the usability of ReMeDy, we implemented an interface to allow quick and efficient data upload by researchers that builds upon the core Signature Commons API. This interface allows data upload using a user friendly CVS file format, while incorporating required data features such as generating universally unique identifier (UUID) links. The upload interface allows for creation of all three data types (pre-clinical and clinical study outcomes as well as assay results for stem cell characterization), and employs metadata validators to define which key value pair elements the metadata must contain, the format of the values (including validation against ontologies), and identify required elements for ingestion. This quality control step allow for final verification of CDE formatting prior to ingestion into the database.

3.2. Multimodular CDE Framework

In order to facilitate data collection and organization within the database, we developed a multimodular CDE framework, which aims to capture all facets of information related to regenerative medicine trials. Previous attempts to create standardized frameworks for characterization of stem cells have resulted in the creation of the Minimum Information About a Cellular Assay for Regenerative Medicine (MIACARM) [4], a format in the process of being adopted by major stem cell banks including hPSCreg. However, it does not provide any CDEs to cover the other areas of clinical or pre-clinical projects needed to gain a full perspective of the work being done. Our multimodular CDE framework aims to address these deficiencies by using a scoping review approach for defining relevant CDEs [5]. The framework consists of 5 modules, which are Project (containing pertinent project status and PI CDEs), Manufacturing/Production (containing donor, Critical Quality Attributes and Critical Process Parameters CDEs), In-depth Product Characterization, Research System (containing study subject and animal model CDEs) and Outcomes. The flexible nature of our modular organization further dovetails with the ReMeDy architecture as not all sections of the framework are required to be shared across all funded projects, allowing for a more flexible and comprehensive organization of all regenerative medicine trials, regardless of their clinical or pre-clinical status, while remaining a comprehensive CDE collection.

3.3. Data Visualization and Sharing

To facilitate future research, collaboration and knowledge generation, ReMeDy utilizes a fully customizable UI and an API for data download. The highly adaptable search function allows to search using both keys and values, as well as incorporating Boolean operators to further refine your query. Further, the use of visualization schemas allows us to tailor visualization of libraries and signatures to highlight the most relevant CDEs to enhance the user experience. The API allows for downloads to be restricted by data type and retrieval of individual data records using the UUID, allowing for flexible, on-demand data retrieval.

4. Discussion

Regenerative medicine is a burgeoning medical field that has only recently been given due attention with the likes of the Regenerative Medicine Innovation Catalyst. Efforts to handle the diverse data generated by regenerative medicine trials, from stem cell characteristics, to clinical and pre-clinical outcomes, to next-generation sequencing big data, are only beginning to get under way. Considering the importance of data management and integration within the regenerative medicine community, here we present the first attempt at integrating stem cell product characterization with patient medical information and clinical outcomes, in order to promote collaboration, foster better standard practices, promote discovery and help drive stem cell product therapies toward FDA approval. The advantage of our architecture is the utilization of a NoSQL approach to import and store data, which allows us to incorporate large data sets in a schema-less nature while simultaneously employing validators to ensure stringent quality control measures and convenient tools for data import and export to improve the long-term usability and viability of our platform.

The next step in evaluating the platform will be further population of the database in coordination with researchers leading the RMIC funded projects. As study subjects are beginning to be enrolled, we plan to continue updating our database on a quarterly basis by assisting researchers in utilizing our upload interface, while simultaneously further improving our access and usability features. Data management and integration, being an important area for healthcare more broadly and regenerative medicine specifically, will only become more critical as this field progresses, a situation which ReMeDy aims to address.

References

[1] Alessio N et al, New Frontiers in Stem Cell Research and Translational Approaches, Biology (Basel) 9(1) (2020), pii: E11.

[2] Keenan AB et al, The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations, Cell Systems 6(1) (2018), 13-24.

[3] Clarke DJB et al, Signature Commons: A Scalable Platform for Serving Diverse Biomedical Data, Submitted for publication.

[4] Sakurai K et al, First Proposal of Minimum Information About a Cellular Assay for Regenerative Medicine, STEM CELLS Translational Medicine 5 (2016), 1345-1361.

[5] Finkelstein J et al, Informatics Approaches for Harmonized Intelligent Integration of Stem Cell Research, Stem Cells and Cloning: Advances and Applications 2020 13 (2020); 1-20.