Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210186

Introducing a Platform for Integrating and Sharing Stem Cell Research Data

Kirill BORZIAK[a,1], Irena PARVANOVA[a], and Joseph FINKELSTEIN[a]
[a] *Icahn School of Medicine at Mount Sinai, New York, New York, USA*

Abstract. Advancements in regenerative medicine have highlighted the need for increased standardization and sharing of stem cell products to help drive these innovative interventions toward public availability and to increase collaboration in the scientific community. Although numerous attempts and numerous databases 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 the platform described in this study, ReMeDy, is to provide 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 relational database. In this initial feasibility study, testing of the ReMeDy functionality was performed using published, publically-available induced pluripotent stem cell projects conducted in in vitro, preclinical and intervention evaluations. It demonstrated the robustness of ReMeDy for storing diverse iPSC data, by seamlessly harmonizing diverse common data elements, and the potential utility of this platform for driving knowledge generation from the aggregation of this shared data. Next steps include increasing the number of curated projects by developing a crowdsourcing framework for data upload and an automated pipeline for metadata abstraction. The database is publically accessible at https://remedy.mssm.edu/.

Keywords. Induced pluripotent stem cell, database, harmonization, common data elements

1. Introduction

Regenerative medicine is a promising therapeutic field, which aims at treatment, repair, and replacing of injured cells, tissues, and organs due to physical damages or degenerative diseases with healthy ones via various mechanisms. This innovative research includes induced pluripotent stem cell (iPSC) therapies advancements that could potentially lead to the successful cure of currently incurable medical conditions. In our previous work, we have discussed the significance of stem cell research, voluminous amount of available stem cell data, and existence of many publically-available stem cell databases [1]. Currently, the existing stem cell data is not consolidated, stored, or available for access by researchers in a centralized and unified manner. Based on the necessity for heterogeneous stem cell data to be organized in a uniform harmonized manner, deposited, and visualized, we created **Re**generative **Me**dicine **D**ata Repositor**y** (ReMeDy) platform [2], which can be publically accessed at https://remedy.mssm.edu/.

[1] Corresponding Author, Kirill Borziak, PhD, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY 10029, USA; E-mail: Kirill.Borziak@mountsinai.org.

ReMeDy is a unique repository, which allows for the systematical collection and sharing of iPSC characteristics, *in vitro* findings and pre-clinical and clinical outcomes, using multimodal common data elements (CDE) framework, allowing for detailed comparisons across studies. Here we present the organization, functionality, and a feasibility study of the ReMeDy platform. We believe that the unique combination of easy accessibility and the diverse range of CDEs stored in ReMeDy has the potential to drive novel knowledge discovery.

In this project we aimed at testing the feasibility of the ReMeDy platform for harmonized curation and storage of data related to diverse iPSC research projects based on the consecutive selection of over 50 published iPSC projects, covering the range *in vitro*, pre-clinical, and clinical studies. The platform feasibility was assessed by its ability to allow each project to be successfully abstracted, stored and accessed using the visualization tools, project links, and Application Programming Interface (API).

2. Materials and methods

2.1. Database architecture and web interface

Our platform, called **Re**generative **Me**dicine **D**ata Repositor**y** (ReMeDy) [2] is an implementation of the Signature Commons, a BD2K-LINCS [3] platform implemented through Docker and designed to store and search diverse metadata in an agile and flexible manner [4]. ReMeDy is installed on a Linux server and implemented through Docker. It is based on the Signature Commons which is available through GitHub at https://github.com/MaayanLab/signature-commons. It is composed of six repositories: controller, data-api, metadata-api, proxy, schema, and ui.

Validation, visualization, and user interface schema were ingested using the API functionality. The API is annotated using Swagger 2.0 JSON implementation, and all RESTful endpoints return structured JSON. Specifically, we developed counting schemas based on the CDE framework that provide additional counting and filtering functionality to the search results page. The schemas, formatted in JSON, were generated and ingested using a custom Python script. To improve usability of the API, the upload process was improved by creating an upload interface. The upload interface was developed using ReactJS for the front-end and Spring Boot on the back-end. The interface allows for uploading and ingestion of CDE templates with minimal command line interface, while maintaining all of the validation features of the default ingestion pipeline.

2.2. Publications used for feasibility study

Data for ReMeDy pilot-testing, in the format of the multi-modular Common Data Elements (CDE) framework [1-2], described below, was obtained from 51 consecutive iPSC articles. We selected the projects based on the criteria that the articles: 1) report iPSCs primary research; 2) represent original research and not reviews; 3) include clinical, preclinical, or in vitro studies; 4) describe cell or Tissue Engineered Medical Products (TEMP). The selected iPSC publications were not restricted by type of a journal, citations, or other criteria. The included projects can be identified by the following PubMed ID numbers: 32632153, 30651323, 32929265, 30535854, 32165680, 25436769, 31445043, 30119058, 24020696, 31227956, 27075820, 28073086, 30582453, 30772682, 32253308, 12084934, 31547869, 32353897, 30691596, 21303266, 26971680, 25143363, 23515118, 30912838, 16308009, 24006477, 33154509, 33142253, 33137106, 33130306, 33108355, 30738321, 28296613, 30224709, 30449714, 31577946, 22495829, 25479750, 22895806, 26494780, 27099175, 28282420, 28436968, 30876823, 31107605, 30442180, 29800782, 31373366, 28967890, 30712489, 23029008.

3. Results

ReMeDy is designed to be a user-friendly database, providing comprehensive and detailed information on induced pluripotent stem cell (iPSC) projects. ReMeDy is freely accessible with no password or registration required. ReMeDy provides a unique resource for collecting, storing, and increasing the utility of data from iPSC publications, as demonstrated by our feasibility study.

3.1. ReMeDy platform

The ReMeDy platform uses a relational database for data storage. Relational databases such as PostgreSQL, which is utilized by ReMeDy, excel at storing and searching structured data in a well-defined schema. Our requirements for well-defined schema, ability to validate against reference ontologies, easily and specificity of searching, and ability to update without compromising data integrity necessitated our selection of a PostegreSQL over a NoSQL approach. The leveraged data technology alongside a properly implemented schema for data storage provides strong data conformity to the FAIR guidelines (Findable, Accessible, Interoperable, and Reusable). Indexing enables very fast searching for nearly any attribute of the metadata without major slowdowns as the size of the tables expand. The ReMeDy platform also contains a stringent metadata validation process, which defines which key value pair elements the metadata will contain, the format of the values (including their validation against ontologies), and identification of required elements for ingestion.

3.2. Multimodular CDE Framework

In order to facilitate data abstraction and organization within the database, we developed a multimodular CDE framework, which aims to capture all facets of information related to iPSC projects. 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 [5], a format in the process of being adopted by major stem cell banks including hPSCreg. However, it does not provide CDEs to cover the full range of information available from published projects. Our multimodular CDE framework aims to address these deficiencies by using a scoping review approach for defining relevant CDEs [1]. The framework consists of 5 modules: Project, Manufacturing / Production, In-depth Product Characterization, Research System and Outcomes.

The Project module contains CDEs which capture information about the PI, general project summary information, such as grant number, project design and regulatory compliance. The Manufacturing / Production module is designed to capture information about the stem cell product under investigation. This includes the source cell information, manufacturing information, and critical final product attributes such as cell marker expression. The In-depth Product Characterization module contains CDEs related to the assays characterizing the stem cell products and source cells, such as transcriptomic profiling. The Research System module contains CDEs related to characterizing study patients, animal models, or *in-vitro* cell lines, as appropriate for each individual project. This module also contains CDEs to characterize any experimental assays the research subjects may undergo during the period of the trial. Finally, the Outcomes / Findings module contains CDEs describing the main findings from the published studies. The modular organization attempts to provide a flexible and comprehensive overview of the potential CDEs that can be used to describe the wide variety of published iPSC projects, and allow for detailed comparisons across studies.

3.3. Data Visualization and Sharing

The landing page is designed to provide easy access to all the functionalities available through the ReMeDy platform. These include the search space functionality, visualization tools, and the API. The search functionality is designed to be maximally function by providing advanced search tools with intuitive controls, which include Boolean operator functionalities, allowing to search both by CDE name and by CDE value. Further improving utility of ReMeDy, we implemented filtering schemas, which serve a dual purpose of incremental refinement of the search query and to provide statistical information on the distribution of CDE values with ReMeDy. To facilitate future research and collaboration, ReMeDy allows researchers to download the database data direction through the fully functioning API. The API provides ample documentation with the aim of promoting easy access and to further foster community sharing and collaboration to help drive advancements is regenerative medicine.

3.4. ReMeDy feasibility testing iPSC project dataset

Results from 51 published clinical and pre-clinical iPSC based studies were used to test the functionality and feasibility of the platform. We abstracted on average 70 CDEs per iPSC study out of a total of 820 CDEs in the multi-modular framework. The feasibility of ReMeDy is demonstrated by the wide diversity of iPSC studies included and successfully abstracted and ingested into the ReMeDy platform. Both autologous and allogeneic treatment types were successfully imported. For patient based studies, testing was successful for clinical, interventional, and observational studies. Both cell products and TEMPs were tested. A wide range of source cell materials were tested including skin, blood, bone marrow, eye, lung, iliac crest, and umbilical cord. Testing pre-clinical studies included studies in mice, rats, pigs, and rhesus macaques. We were also able to abstract a wide range of disease conditions, with 15 different conditions documented in the database, including cancer, age-related macular degeneration, amyotrophic lateral sclerosis, Gaucher disease and graft versus host disease.

4. Discussion

The expanding field of regenerative medicine required the creating of a flexible and agile repository for stem cell data aggregation, storage, visualization, and sharing; therefore, we launched **Re**generative **Me**dicine **D**ata Repositor**y** (ReMeDy) platform. ReMedy is an organized repository which captures iPSC-research project information in standardized and effortless to visualize format. The advantages of our architecture allows us to incorporate large data sets while employing validators to ensure stringent quality control measures. The platform includes a flexible common data elements framework, which allows harmonization and standardization of all currently and future stored data. It promotes easy accessibility to stem cell projects to facilitate data sharing and collaboration within the field. The primary utility of ReMeDy over PubMed and ClinicalTrials.gov is the ability to quickly obtain information and statistics on a broad range of CDEs and their interrelations, due to the diverse data which was abstracted into the multi-modular CDE framework. For example, ReMeDy is able to provide information on which animal models a particular iPSC line has been tested and as a treatment for which diseases/conditions with a single search.

We tested the feasibility of the ReMeDy platform to store published stem cell projects, including *in vitro*, pre-clinical, and clinical studies, in a systemic organized manner, using flexible framework of CDEs by uploading an initial set of 51 PubMed-indexed projects. Our feasibility study illustrates the simplicity and flexibility of ReMeDy, allowing for the storage and visualization of metadata from any stem cell project, when utilizing the standardized multi-modular CDE framework as a template for data acquisition. Future plans for ReMeDy include updates and automation of the platform, by implementing crowdsourcing and natural language processing, using MeSH terminology and ontology-driven functionalities [6]. The application of these approaches will allow us to greatly expand the number of publications available in ReMeDy, realizing the potential of driving knowledge discovery through the use of statistical and comparative analyses of the abundance of iPSC data that will be available in ReMeDy.

5. Conclusion

ReMeDy database allows consolidation, storage, and availability for access by researchers in a centralized and unified manner. Our current feasibility analysis shows that the ReMeDy platform provides a practical solution to the harmonization of diverse iPSC data across a range of publication into a single multi-modular framework. The advantages provided by easy access to aggregated CDE values and statistics has the potential to drive knowledge discovery.

Reference

[1] Finkelstein J, Parvanova I, Zhang F. Informatics Approaches for Harmonized Intelligent Integration of Stem Cell Research. Stem Cells Cloning. 2020 Jan 28;13:1-20.

[2] Borziak K, Qi T, Evangelista JE, Clarke DJB, Ma'ayan A, Finkelstein J. Towards Intelligent Integration and Sharing of Stem Cell Research Data. Stud Health Technol Inform. 2020 Jun 26;272:334-337.

[3] Keenan AB, Jenkins SL, Jagodnik KM, et al. The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. Cell Syst. 2018 Jan 24;6(1):13-24.

[4] Stathias S, Turner J, Koleti A, et al. LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. Nucleic Acids Research. 2020 48(D1):D431–D439.

[5] Sakurai K, Kurtz A, Stacey G, Sheldon M, Fujibuchi W. First Proposal of Minimum Information About a Cellular Assay for Regenerative Medicine. Stem Cells Transl Med. 2016 Oct;5(10):1345-1361.

[6] Elghafari A, Finkelstein J. Introducing an Ontology-Driven Pipeline for the Identification of Common Data Elements. Stud Health Technol Inform. 2020 272:379-382.