*Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era*
*J. Delgado et al. (Eds.)*


*doi:10.3233/SHTI210826*

# Intelligent Integrative Platform for Sharing Heterogenuous 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.** Recent studies demonstrated that comparative analysis of stem cell research data sets originating from multiple studies can produce new information and help with hypotheses generation. Effective approaches for incorporating multiple diverse heterogeneous data sets collected from stem cell projects into a harmonized project-based framework have been lacking. Here, we provide an intelligent informatics solution for integrating comprehensive characterizations of stem cells with research subject and project outcome information. Our platform is the first to seamlessly integrate information from iPSCs and cancer stem cell research into a single platform, using a multi-modular common data element framework. Heterogeneous data is validated using predefined ontologies and stored in a relational database, to ensure data quality and ease of access. Testing was performed using 103 published, publicly-available iPSC and cancer stem cell projects conducted in clinical, preclinical and in vitro evaluations. We validated the robustness of the platform, by seamlessly harmonizing diverse data elements, and demonstrated its potential for knowledge generation through the aggregation and harmonization of data. Future aims of this project include increasing the database size using crowdsourcing and natural language processing functionalities. The platform is publicly available at https://remedy.mssm.edu/.

**Keywords.** Common data elements, induced pluripotent stem cells, cancer stem cells.

## 1. Introduction

Stem cells were first described in 1961 by James Till and Ernest McCulloch [1]. Today, stem cell research has dramatically transformed and advanced the field of regenerative medicine. Due to the large number of published stem cell research studies, researchers aim to collect, store, and centralize the gathered data. In previous publications by our team, we have developed and tested Regenerative Medicine Data Repository (ReMeDy) platform, allowing collection and sharing of *in vitro* findings and pre-clinical/ clinical trial outcomes [2, 3]. Currently, our platform contains 103 stem cells research papers, included in the PubMed database. Each featured project can be accessed across the framework by utilization of user-friendly tools and API platforms, due to the use multi-modal flexible common data elements (CDE) framework, which permits cross-studies comparison and collaboration.

[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.

## 2. Methods

### *2.1. Database architecture and web interface*

Our platform, **Re**generative **Me**dicine **D**ata Repositor**y** (ReMeDy) [1], is an implementation of the Signature Commons (https://github.com/MaayanLab/signature-commons), which is a BD2K-LINCS DCIC platform [2], installed through Docker and designed to store and search diverse metadata in an agile and flexible manner [4]. The ReMeDy platform was installed using the default instructions on a Linux server. It contains six repositories: controller, data-api, metadata-api, proxy, schema, and ui.

The various validation, visualization, and user interface schema were ingested through the Application programming interface (API) functionality. Specifically, we developed counting schemas based on the CDE framework, which aim to 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 the utility of the API, we developed an upload interface, which automated the ingestion process. The upload interface was developed using ReactJS and Spring Boot. The interface allows for uploading and ingestion of CDE templates without command line interface, while maintaining the validation features.

### *2.2. Literature search and data abstraction*

To test the ability of ReMeDy to handle heterogeneous stem cell data, we selected a set of 103 iPSC and CSC original research publications, using a randomized process from Google Scholar and PubMed search results for “iPSC” and “cancer stem cells”, respectively. The randomized selection process was designed to ensure the inclusion of the full range of stem cell research. Further, we ensured the inclusion of in vitro, pre-clinical/animal model, and clinical trials of iPSC and CSC publications.

Following the selection of our publication set, the data from the publications was abstracted into the multi-modular Common Data Elements (CDE) framework [2, 4]. The abstraction process was conducted manually by trained abstractors with experience in cancer, regenerative medicine, and stem cell research. The majority of abstracted CDE values were defined either by permissible value sets or by ontologies. CDEs which are not amenable to being extracted as specific values, such as outcomes and findings descriptions, were recorded as short statements in free-text value fields. Further, a template was created for each cell line, individual, or grouped study subjects. The templates were then submitted to the upload interface utility, converted to JSON, ingested, and validated trough the API [5].

## 3. Results

The ReMeDy platform is a user-friendly database, which contains comprehensive and detailed information from stem cell research publications. The focus of the current iteration of ReMeDy is to seamlessly integrate induced pluripotent stem cell (iPSC) and cancer stem cell (CSC) projects. ReMeDy is currently freely accessible with no registration requirements.

### 3.1. The ReMeDy platform

The ReMeDy platform takes advantage of a relational database for data storage, such as PostgreSQL, which is implemented in our platform, excel at storing and searching structured data through organizing data within a well-defined schema (Figure 1). With the aim to conform to the FAIR guidelines (Findable, Accessible, Interoperable, and Reusable), our requirements for well-defined schema, validation against reference ontologies, ease and specificity of searching, and the ability to update data without compromising its integrity drove us to select it over a NoSQL approach. Further, indexing of the data enables for very fast searching of any attribute of the metadata without major slowdowns as the size of the tables expand. Our stringent metadata validation process includes strict definitions of key value pairs, the proper formatting of the values, and specification of required elements.

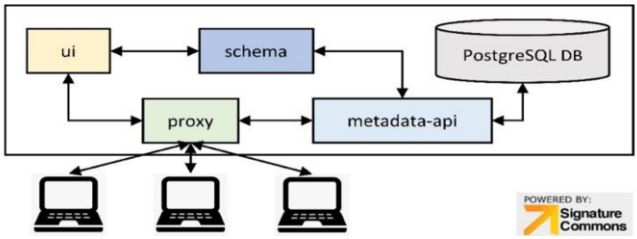


**Figure 1.** ReMeDy platform architecture displaying the interconnection of the Signature Commons packages.

### 3.2. Multi-modular CDE framework

In order to promote data harmonization and to facilitate data abstraction, we developed the multi-modular CDE framework. Our aim was to capture all the various facets of information related to iPSC and CSC projects. Previously standardized frameworks for characterization of stem cells, such as the Minimum Information About a Cellular Assay for Regenerative Medicine [6], do not cover the full range of information available from published projects and are limited to stem cell features and assays used to derive them. Our multi-modular CDE framework addresses these deficiencies by using a scoping review approach for defining relevant stem cell characteristic-related CDEs [7]. The resulting framework consists of 5 modules: Project, Stem Cell Characteristics, In-depth Characterization, Research System, and Outcomes / Findings (Figure 2).

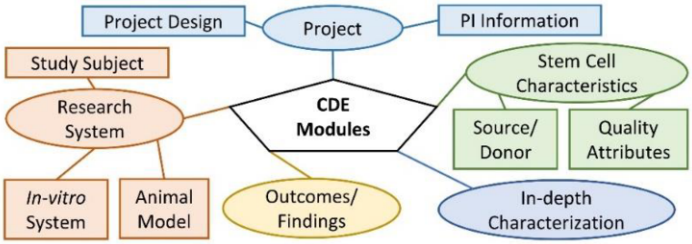


**Figure 2.** Schematic of the multi-modular CDE template, highlighting the modules and their CDE content.

The Project module CDEs capture general project information, such as PI contact information, funding information, publication information, and project design. The Stem Cell Characteristics module is designed to capture information about the stem cell

products under investigation. The In-depth Characterization module contains CDEs related to different assays that can be used to characterize the stem cell, such as transcriptomic profiling, clonal capacity, or genetic stability. The Research System module CDEs characterize the study patients, animal models, and/or *in-vitro* cell lines. Finally, the Outcomes / Findings module CDEs describe the outcomes of clinical studies and findings from pre-clinical studies. Since not all CDEs are required for all studies, our modular organization provides a flexible approach for comparisons across studies.

### *3.3. Data accessibility, visualization and sharing*

The ReMeDy site provides easy access to the various functionalities, such as search functionality, visualization tools, and API. It allows a search by CDE name or CDE value. Further, implemented filtering schemas allow users to incremental refinement of their search queries, and provide statistical information on the distribution of CDE values among the ReMeDy projects. ReMeDy also allows researchers to download the abstracter data directly through the API with the aim of promoting easy access, community sharing, and collaboration to advance stem cell research.

### *3.4. ReMeDy feasibility testing*

To test the functionality and feasibility of our platform, we used 103 published clinical, pre-clinical, and *in vitro* iPSC and CSC studies. We abstracted on average 76 CDEs per study of total of 841 CDEs comprising the multi-modular framework. ReMeDy's feasibility was demonstrated by diversity of publications from the US, China, Japan, and Italy, amongst others. Abstraction of a wide range of source cell materials was tested (skin, blood, bone marrow, and others). Pre-clinical studies included studies in mice, rats, pigs, and rhesus macaques. We were able to abstract 15 different disease conditions, including cancer, heart disease, sclerosis, spinal cord injury, and others (Figure 3).

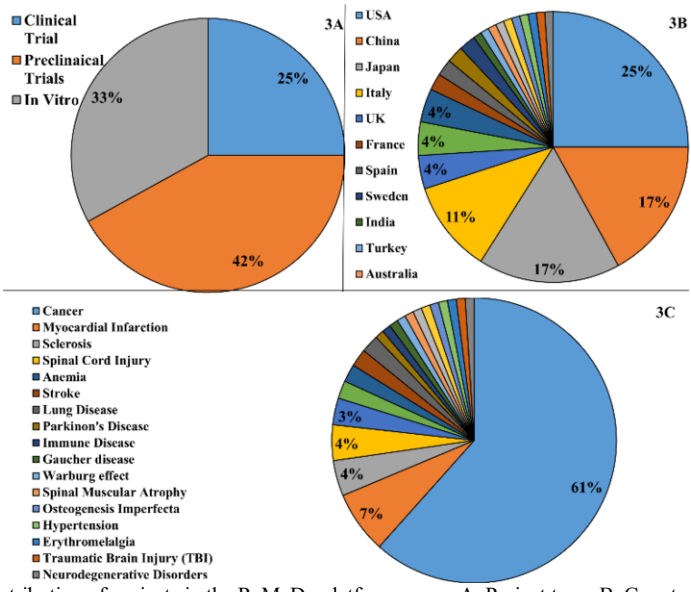


**Figure 3.** Distribution of projects in the ReMeDy platform across A. Project type; B. Country, conducting the research; C. Cancer type.

## 4. Discussion

The expanding field of stem cell research in both regenerative and cancer medicine requires the creation of a flexible and agile repository for data aggregation, storage, visualization, and sharing. To promote this effort, we have adapted the Regenerative Medicine Data Repository (ReMeDy) platform and the multi-modular CDE framework for use with both iPSC and CSC projects. A primary advantages of ReMeDy is its organized multi-modular framework, which harmoniously captures both iPSC and CSC research project information in a standardized format and provides effortless visualization. The platform was tested by uploading 103 clinical, preclinical, and in vitro studies, in a systemic manner, confirming ReMeDy to be a harmonized storage and visualization platform for diverse stem cell data. The relational JSON formatted database allows us to import CDE data, while employing validators for a stringent quality control.

Future aims for ReMeDy include increasing the database size to include all published iPSC and CSC research. This will be accomplished by implementing natural language processing and crowdsourcing functionalities. To automate data abstraction, we aim to use MeSH terminology and ontology-driven functionalities [8, 9]. These approaches will allow us to realize the potential of driving knowledge discovery through the use of statistical and comparative analyses of iPSC and CSC data. Crowdsourcing functionality will be implemented by expanding our iPS and CSC automated pipeline.

## 5. Conclusion

The ReMeDy platform allows for consolidation, harmonization, and storage of diverse stem cell CDEs, available for access in a centralized and unified manner. The platform provides the first attempt to abstract iPSC and CSC data into a single unified framework. The access to and analysis of harmonized CDEs has the potential for generation of new knowledge and advance regenerative and cancer medicine.

## References

[1] Biehl JK, Russell B. Introduction to stem cell therapy. J Cardiovasc Nurs. 2009;24(2):98–105.
[2] Borziak K, Parvanova I, Finkelstein J. ReMeDy: a platform for integrating and sharing published stem cell research data with a focus on iPSC trials. Database (Oxford). 2021;2021.
[3] Parvanova I, Borziak K, Finkelstein J. A Platform for Integrating and Sharing Cancer Stem Cell Data. Annu Int Conf IEEE Eng Med Biol Soc. 2021.
[4] Stathias V, Turner J, Koleti A, et al. LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. Nucleic Acids Res. 2020;48(D1):D431-d9.
[5] Borziak K, Qi T, Evangelista JE, et al. Towards Intelligent Integration and Sharing of Stem Cell Research Data. Stud Health Technol Inform. 2020;272:334-7.
[6] Sakurai K, Kurtz A, Stacey G, et al. First Proposal of Minimum Information About a Cellular Assay for Regenerative Medicine. Stem Cells Transl Med. 2016;5(10):1345-61.
[7] Finkelstein J, Parvanova I, Zhang F. Informatics Approaches for Harmonized Intelligent Integration of Stem Cell Research. Stem Cells Cloning. 2020;13:1-20.
[8] Elghafari A, Finkelstein J. Introducing an Ontology-Driven Pipeline for the Identification of Common Data Elements. Stud Health Technol Inform. 2020;272:379-82.
[9] Elghafari A, Finkelstein J. Automated Identification of Common Disease-Specific Outcomes for Comparative Effectiveness Research Using ClinicalTrials.gov: Algorithm Development and Validation Study. JMIR Med Inform. 2021;9(2):e18298.