INFORMATICS PROFESSIONALS. LEADING THE WAY.

Regenerative Medicine Data Repository (ReMeDy) for Harmonized Aggregation and Sharing of Stem Cell Research Data

Panel - Intelligent Integrative Informatics Approaches for Big Data Aggregation, Sharing and Analytics in Stem Cell Research

S15

Kirill Borziak, PhD, John E. Evangelista, MS, Daniel Clarke, MS, Avi Ma'ayan, PhD, Joseph Finkelstein, MD, PhD

Icahn School of Medicine at Mount Sinai, New York, NY

Disclosure

I have no relevant relationships with commercial interests to disclose.

Learning Objectives

After participating in this session the learner should be better able to:

- Learn major building principles of Regenerative Medicine Data Repository (ReMeDy)

ReMeDy platform

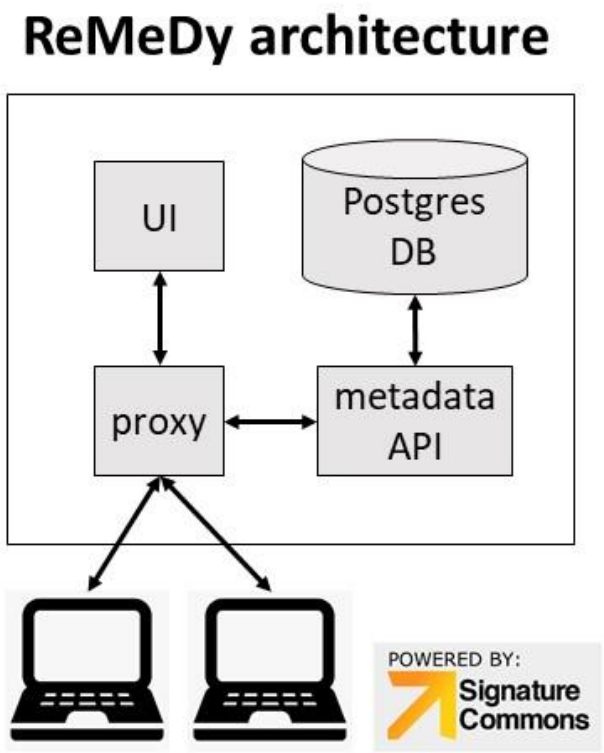
The function of ReMeDy is to provide storage and dissemination services for in-depth stem cell clinical products characterization results, clinical data, and clinical outcomes that are generated as part of regenerative medicine research.
ReMeDy is an implementation of the Signature Commons.

Signature Commons was designed as part of NIH-funded BD2K-LINCS Data Coordination and Integration Center to store and search diverse metadata in an agile and flexible manner.
Signature Commons was developed with the aim of maintaining data access and interoperability, persistent identifiers, data citation, and data reuse. The goal is that the data is easily accessible and consistent with the findable, accessible, interoperable, and reusable (FAIR) principles.

Relational database

- Signature Commons utilizes PostgreSQL relational database for data storage to ensure FAIR principles.
- PostgreSQL was utilized to allow easy updating, indexing, and fast search times.
- PostgreSQL ensures the integrity of the data by placing constraints based on the definition of primary and foreign keys.
- PostgreSQL allows us to maintain the relationships of the highly structured data, implemented through multi-modular CDE framework.
- PostgreSQL ensures that the data is easily and specifically searchable, and is updatable without risking to compromise data integrity.
- Indexing enables very fast searching of the metadata without experiencing slowdowns as the volume of regenerative medicine data expands.

ReMeDy architecture



- Signature Commons is installed through Docker and consists of six packages: controller, data-api, metadata-api, proxy, schema, ui.
- The main functions of the database, are done by the proxy, metadata-api, and ui.

ReMeDy architecture

- The data-api repository is responsible for the set up and installation of the ReMeDy instance.
- It is built using Gradle and Java.
- data-api compiles production Java source files using the JDK compiler, deploy Web Application Resource (WAR) files, performs verification tasks such as running the tests, and run the unit tests with Tomcat using JUnit and TestNG.

ReMeDy architecture

- The proxy repository is used as the primary access point to the ReMeDy platform and provides the coordination between the user interface (UI) and API services.
- It is a convenient and generic proxy image, built in nginx.
- It functions to provide a single ingress for the multiple microservices of ReMeDy that is configurable via environment variables.

ReMeDy architecture

- The ui repository is the front-end UI for displaying the stem cell data.
- It is built mainly in javascript.
- This repository is responsible for the visualization of stored data, the search functionalities, and graphical representation tools.

- The metadata-api repository is responsible for communication with the PostgreSQL DB.
- It is built using LoopBack and TypeScript.
- This is the primary API implementation that functions both to retrieve and upload metadata into ReMeDy.

ReMeDy architecture

- The controller repository is an intermediary to aid the data ingestion process.
- controller is built with Python and Java.
- This repository include scripts to facilitate metadata ingestion from different forms into JSON which is use by ReMeDy, and interacts with the API to facilitate upload.

- The schema package contains information on the JSON-Schema validators for ReMeDy entities.
- The schema files are in JSON format, with testing performed using TypeScript.
- The schema are designed to flexibly validate arbitrary metadata in the ReMeDy database.

ReMeDy API

- The comprehensive API that powers ReMeDy allows users to find and access the data programmatically.
- The API is powered by LoopBack and is responsible for communication with the PostgreSQL DB for both retrieval and uploading of data.
- It is designed with a Swagger 2.0 JSON implementation, with all the RESTful endpoints returning structured JSON format data.
- The multiple endpoints allow for searching of the content by any of the annotations, data value and universally unique identifiers (UUIDs).
- Different slices of metadata can also be requested that are associated with desired subcategories of stem cell data stored in ReMeDy.

ReMeDy schemas

```
{
  "$id": "/dcic/signature-commons-schema/v5/meta/resource/sigcom.json",
  "$schema": "http://json-schema.org/draft-04/schema#",
  "description": "A sample validator",
  "allOf": [
    {
      "$ref": "/dcic/signature-commons-schema/v5/core/meta.json"
    },
    {
      "properties": {
        "Name": {
          "description": "The name of this entry",
          "type": "string"
        },
        "URL": {
          "description": "URL of the entry",
          "type": "string",
          "format": "uri"
        },
        "Created": {
          "description": "Date created",
          "type": "string",
          "format": "date"
        },
        "Role": {
          "description": "Role of entry",
          "type": "string",
          "enum": ["admin", "user", "guest"]
        },
        "Value": {
          "description": "Some value",
          "type": "number",
          "minimum": 0,
          "maximum": 5
        }
      },
      "required": [
        "Name",
        "Value"
      ],
      "type": "object"
    }
  ]
}
```

- ReMeDy utilizes JSON format to ingest schemas and data.
- Advantages of JSON for structured data over XML is that it uses less data overall, increases parsing speed, and increases readability of retrieved data.
- Using JSON further allows for implementation of schemas, which are utilized to validate data prior to ingestion.

ReMeDy data validation

- To ensure data quality prior to final ingestion of the stem cell data into ReMeDy, all input data is subjected to validation against predefined value sets and existing ontologies that ensure proper formatting and interoperability of the stem cell data.
- This includes, for example the validation of protein names against the UniPROT standard and gene names against the NCBI Gene standard.

ReMeDy data organization

Resource:
- Master container for each published stem cell project
- Contains libraries

Library:
- Container for project CDEs, e.g. Project, Manufacturing / Production Modules, Outcomes
- Contains signatures

Signature:
- Represents individual/ grouped Research System data
- Contains information for patients, animal models, and in vitro subjects

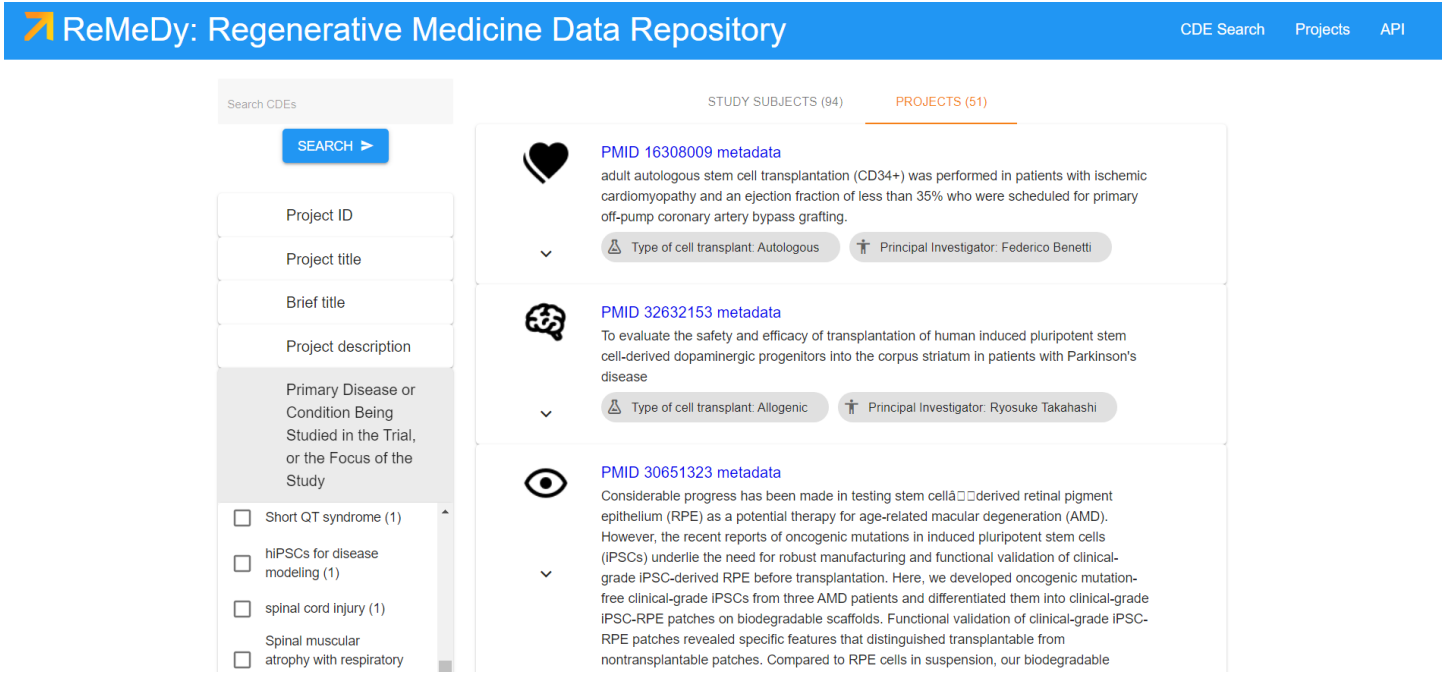
Validator:
- Quality control file
- Used to verify formatting and requirements of JSON files, prior to ingestion
- Allows for integration of ontology information
- Applicable to resources, libraries and signatures

Schema:
- Visualization control file
- Defines how the data is presented in the search results
- Applicable to resources, libraries and signatures

The data contained in ReMeDy is hierarchically organized into three layers:

- resources are the top level containers that represent individual research projects,
- libraries contain the data elements that characterize distinct formulations the stem cell products,
- signatures contain data that characterize the patients and animal models which were treated with the stem cell products.

ReMeDy data access features

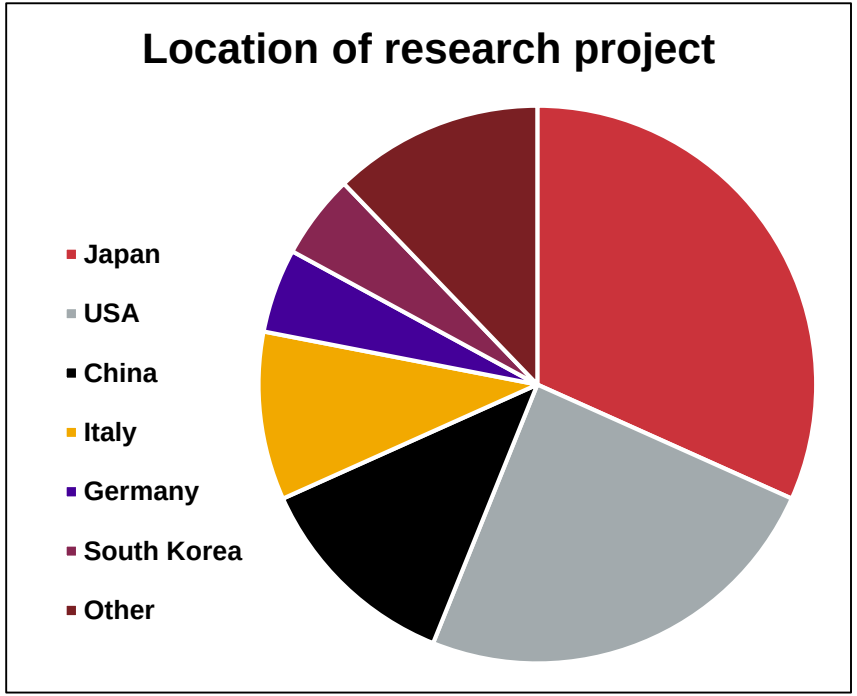


Advanced search interface and filtering schemas

ReMeDy feasibility testing

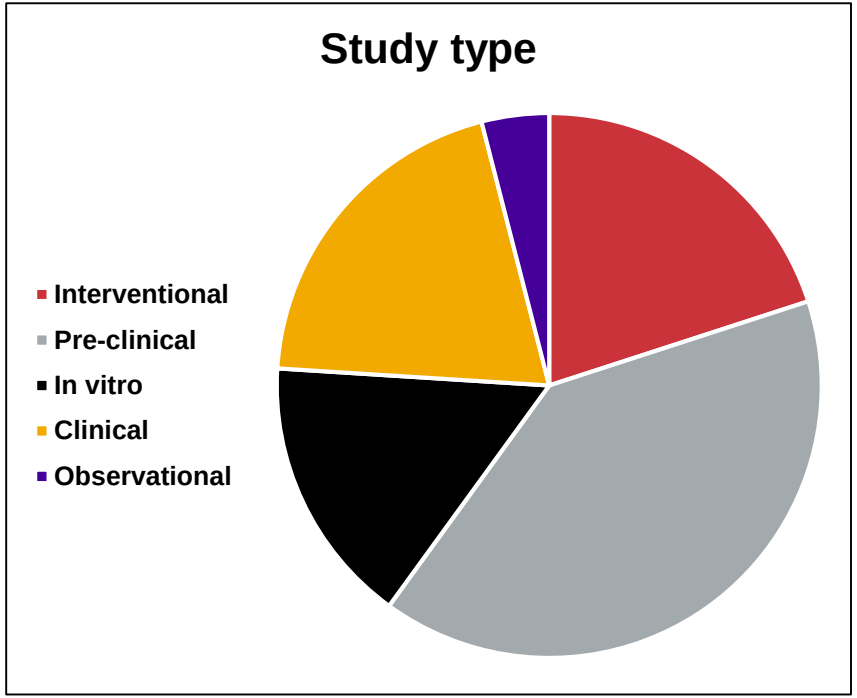
- To test the feasibility of ReMeDy in storing stem cell data, we testing the platform with randomized set of 51 published induced pluripotent stem cell projects, indexed in PubMed.
- Stem cell metadata from the publications was abstracted into the multi-modular CDE framework, by trained abstractors.
- All resulting data files were successfully validated and ingested into ReMeDy, using the ReMeDy schema and API frameworks.

ReMeDy feasibility testing



Our feasibility testing covered data from 11 countries: Japan, USA, China, Italy, Germany, South Korea, Sweden, Argentina, Saudi Arabia, UK, and Pakistan.

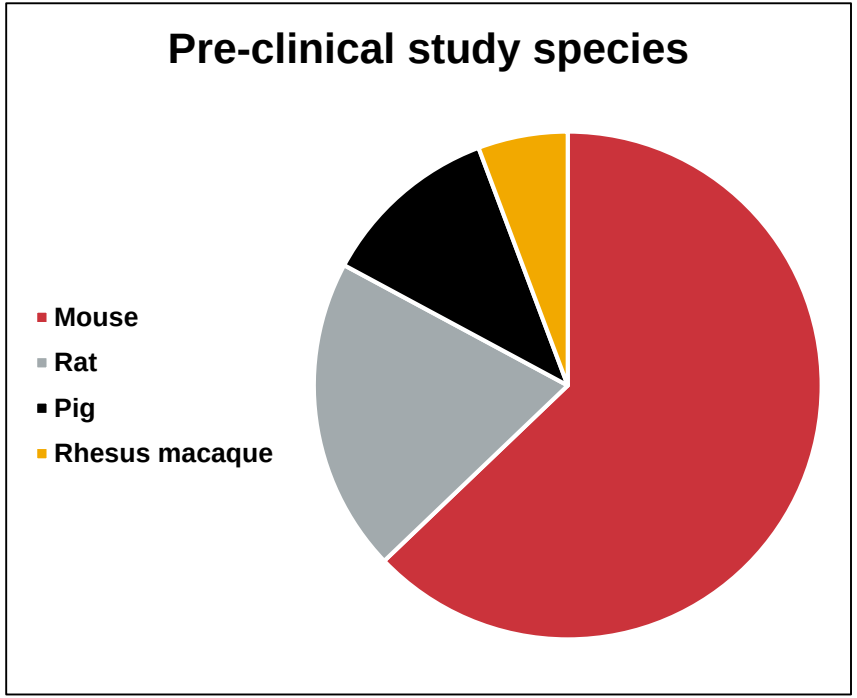
ReMeDy feasibility testing



ReMeDy is able to accommodate clinical, pre-clinical and, in vitro studies.

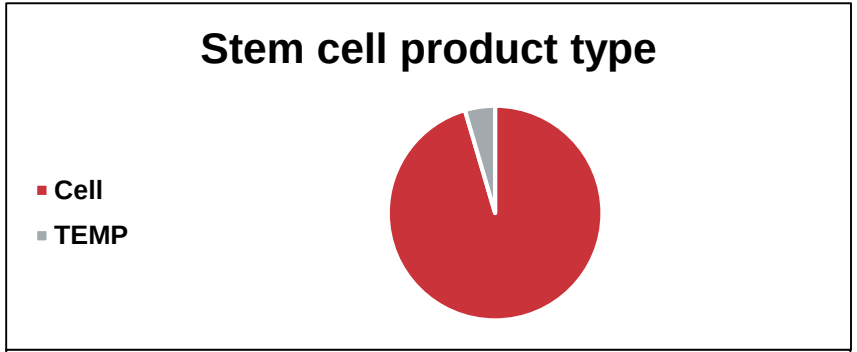
Among clinical studies, we were able to abstract and ingest both interventional and observational studies.

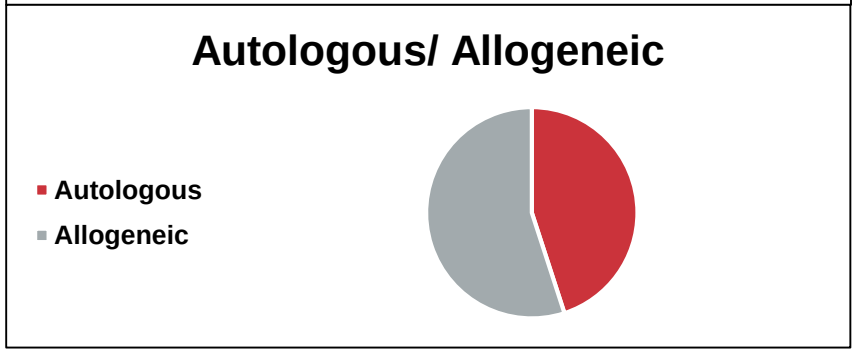
ReMeDy feasibility testing



ReMeDy is able to accommodate a diverse range of animal models by adapting the multi-modular CDE framework.

ReMeDy feasibility testing

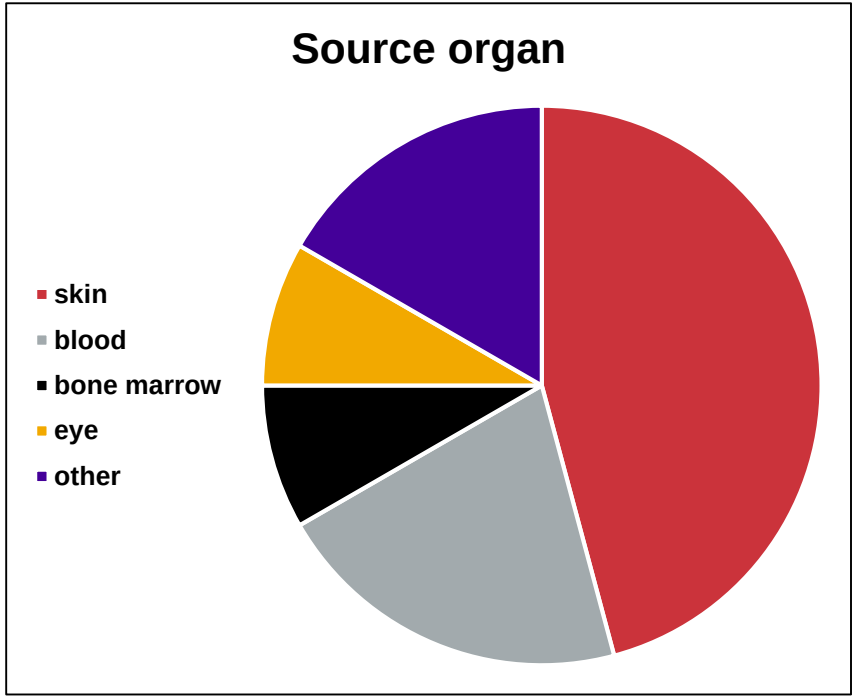




ReMeDy is able to accommodate both stem cell studies and Tissue Engineered Medical Product studies.

We were able to successfully accommodate for autologous and allogeneic stem cell studies.

ReMeDy feasibility testing



The tested projected also covered a diverse range of source tissues and protocols for iPSC generation.

ReMeDy feasibility testing

A diverse range of primary disease conditions are able to be successfully abstracted into ReMeDy, including:

- cancer
- age-related macular degeneration
- hamyotrophic lateral sclerosis
- aplastic anemia complicated by platelet transfusion refractoriness
- end-stage respiratory malfunctions
- gaucher disease
- graft versus host disease
- ischemic cardiomyopathy
- multiple sclerosis
- Neurofibromatosis type 1
- Parkinson's disease
- retinitis pigmentosa
- spinal cord injury

ReMeDy availability

ReMeDy can be accessed at:
https://remedy.mssm.edu/

Signature Commons is available at:
https://github.com/MaayanLab/signature-commons

Thank you!

Email me at: Kirill.Borziak@mountsinai.org