HUMAN PLURIPOTENT STEM CELL REGISTRY

A Data Management Plan created using DMPonline

Creator: Nancy Mah

Affiliation: Other

Template: European Commission (Horizon 2020)

ORCID iD: 0000-0003-3301-6546

Project abstract:
Human pluripotent stem cells (hPSC) can differentiate into all cell types of the human body and can be expanded without limit. The cells are genetic surrogates of their donors. These three characteristics (potency, expandability, personalization) make these cells enormously valuable for cell therapy and regenerative medicine, for disease and tissue modeling and for testing of drugs, toxins and chemicals. While human embryonic stem cell (hESC) lines are derived from early human embryos, induced pluripotent stem cells can be generated from any somatic cell. Both cell types are currently being used in research, including in clinical trials. To have knowledge about available hPSC ines, their quality, their ethical provenance and availability, a central registry is required to provide this information. The human pluripotent stem cell registry fulfills this task. It is the most accepted and complete hPSC-registry providing transparency in the field, an invaluable resource of cell lines and their application and developing and promoting standards and guidelines. In addition, the Registry reduces risks in the field by enabling comparability and reproducibility, as well as avoiding waste of resources by providing access to hPSC-lines worldwide. The Registry is also instrumental in promoting quality standards of lines. Substandard lines will not be accepted for use by the community, while lines in hPSCreg fulfill a validated standard. Hence, the hPSCreg is a global hub for hPSC data and information. The objectives of the project are driven by the needs for a comprehensive public information source for pluripotent stem cells as stated in the H2020-EU.3.1 call (Societal challenges - Health, demographic change and well-being): to gather and make available detailed information on the different hPSC lines derived in Europe and beyond, thereby also avoiding needless creation of new cell lines. This registry operates through an internet website that will continue to provide high quality data about the lines (e.g. cell characteristics), details regarding their source and contact information regarding their location. The objectives are divided into the main concept areas (i) acquisition, registration and qualification of information, (ii) communication, dissemination and harmonization, (iii) technical implementation, governance and project management. The specific objectives are (1) Provision of validated ethics information for each cell line, (2) Provision of validated scientific information for each cell line, (3) Provision of information on clinical application of hPSC lines, (4) Establish and strengthen international reach and community interaction as well as acceptance, (5) A stable state-of-the-art back- and front-end must be established and maintained and (6) Implement a synergistic management structure supported by clear governance tasks.

Last modified: 2019-11-05

HUMAN PLURIPOTENT STEM CELL REGISTRY - DETAILED DMP

1. DATA SUMMARY

State the purpose of the data collection/generation

The Human Pluripotent Stem Cell Registry (hPSCreg; https://hpscreg.eu/) was originally founded to impart transparency in the use of human pluripotent stem cell lines in European Union-funded research. To this end, hPSCreg collects data about the ethical provenance and biological properties of human pluripotent stem cell lines. Based on this user-provided data, hPSCreg issues certificates for cell lines that fulfill ethical and scientific standards as set out by the European Commission and the stem cell community, including members of the Committee of National Representatives from EU countries. Although its roots are in the EU, hPSCreg has expanded to become an international registry of human pluripotent stem cell lines, and through the collection of data using a standard field structure and ontologies, enables the comparison and evaluation of hPSC lines generated in different countries and at multiple sites, including core institute cell supply facilities, individual research laboratories, and biobanks.

Explain the relation to the objectives of the project

As a registry for human pluripotent stem cell lines, the collection of stem cell line data is the primary objective of the project.

Specify the types and formats of data generated/collected

Technical details. Data are collected using a web-based interface. Most of the data is collected as text, either as limited selections from a drop-down menu, clickable boxes, or free text. APIs or internal software are used to generate standardised or interchangeable data formats for BioSamples IDs (BioSamples API from EMBL-EBI; https://www.ebi.ac.uk/biosamples/docs/references/api), the generation of the standard cell line name (stem cell community standard; https://dx.doi.org/10.1016%2Fj.stemcr.2017.12.002), and cell types and clinical phenotypes (ZOOMA text-to-ontology mapping tool from EMBL-EBI; https://www.ebi.ac.uk/spot/zooma/). For some data fields, supplementary information is accepted as documents (e.g. *.docx, *.pdf) or image files (e.g. *.tiff, *.png, *.jpg). The change history is stored in git repositories. All other data is stored in a relational database (MySQL) and in a text-based database (Elasticsearch; https://www.elastic.co/de/products/elasticsearch).

Data description. A variety of data are collected by hPSCreg: 1) user data; 2) cell line data; 3) project data; 4) clinical study data.

1. **User data**: cell lines may only be entered by registered users of hPSCreg. To this end, a user must register at hPSCreg with contact details, including full name, institution, email and address. hPSCreg also tracks user activity on the website using the Open Source Tool Matomo (https://matomo.org/) for the purposes of optimizing the function of the Registry. Tracked data is used internally to enrich the user experience and no tracked data is given to third parties.

   hPSCreg keeps a history of changes made to cell lines and other objects and who made them. This provenance makes it possible to check where inconsistencies or other problems are introduced and helps to make corrections.

   We keep standard server logs (including IP addresses) to track down problems with the server. These logs are used in anonymized form to generate usage statistics of the website. The unanonymized data is removed by automated cleanup after at most a couple of months.

2. **Cell line data**: the Registry collects over one thousand fields of data relating to the ethical provenance and biological properties of the cell lines. Briefly, this includes details about the donor of the material used to derive the hPSC lines, including the conditions of donor consent, which have a bearing on the downstream usage of the hPSC line derived from donor material, and clinical data about the donor such as age and disease phenotypes. Data on the biological properties of the cell line, including the derivation of the line, culture conditions, evidence of pluripotency and genetic constitution of the cell line (for example, genotyping data, STR profiles and genetic modifications) are also recorded in the Registry. Sensitive genetic data such as STR/ HLA profiles and sequencing data are either stored directly in hPSCreg or in a public repository (such as the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/home)), respectively. Access to these sensitive data is subject to the approval of the hPSCreg Data Access Committee.

3. **Project data**: users have the option of linking hPSCreg-registered lines to public or privately funded projects. Project data consists of a title and short description, project funder, funding period and contact institution, as well as hPSC lines associated with the project.

4. **Clinical study data**: hPSCreg maintains a registry of clinical studies involving hPSC lines for cell therapy. Clinical study data may be entered by registered users or the hPSCreg administration team. Data from clinical studies includes basic information about clinical trials, including title, description, regulatory authority and clinical trial identifier, study contact and type of hPSC line and its derived cell type for therapy, but to name a few fields.

Specify if existing data is being re-used (if any)

Data on hPSCreg that originates from the re-use of data could include data that has already been published in scientific journals, such as figures. Users who enter the cell line data into hPSCreg are responsible for making sure that the data they enter into hPSCreg does not violate any original copyrights on the data. Other re-used data in hPSCreg include data from public clinical trial registries, such as the WHO International Clinical Trials Registry Platform (http://apps.who.int/trialsearch/). According to the hPSCreg Terms of Use (https://hpscreg.eu/terms), the Registry places no additional constraints on data re-use.

Specify the origin of the data

The data in hPSCreg is entered by registered users of hPSCreg or the hPSCreg Administration Team. Cell line data originates from user-provided data, or in the case of cell line data entered by hPSCreg Administration Team, from publications in peer-reviewed journals. Project data originates from the user, and this data may already be publicized through a project website or entry in a project database such as CORDIS (https://cordis.europa.eu/projects/en). Data about clinical studies originates primarily from public clinical trial registries, such as ICTRP 8https://www.who.int/ictrp/en/), and this data may be supplemented by direct communications between hPSCreg and the clinical study contact.

State the expected size of the data (if known)

The current total size of hPSCreg is about 6.5 GB (4500 cell lines), with the bulk of the space taken up by uploads associated with the cell lines (e.g. pdf and image files). The git repository with the change history takes ca. 400 MB, The cell line data itself occupies about 23 MB in MySQL and 28 MB in Elasticsearch. It is projected that hPSCreg will expand to over 10000 lines by the end of 2020, requiring approximately 20 GB of storage space (assuming ~2 MB for one cell line).

Outline the data utility: to whom will it be useful

The generation of hPSC lines at multiple sites, such as stem cell core facilities, individual research laboratories, biobanks, etc., inevitably leads to a high degree of variability in the availability of donor information and characterisation and production process generating hPSC lines. The Registry provides a means to evaluate hPSC lines on their ethical provenance and biological properties, through the standardised collection of pluripotent stem cell line data. This information is useful to all stakeholders as follows:

- **Academia**: hPSC lines with specific properties (e.g. disease context) can be found

- **Industry**: hPSC lines with favourable licencing conditions for commercial applications can be found

- **Regulators and Funders**: the data collection serves to provide an overview of research outputs from EU-funded projects that involve hPSC lines. These outputs may include additional hPSC lines, publications, continued use in other EU-funded projects, or clinical translation in the form of clinical grade hPSC lines for cell therapy.

2.1 MAKING DATA FINDABLE, INCLUDING PROVISIONS FOR METADATA [FAIR DATA]

Outline the discoverability of data (metadata provision)

Registered lines in hPSCreg all have been assigned a unique standard cell line name according to a stem cell community standard nomenclature (https://dx.doi.org/10.1016%2Fj.stemcr.2017.12.002). These standard names can be used in publications (e.g. Stem Cell Research - Lab Resource) and other research outputs to

unequivocally identify the lines. Additionally, hPSCreg lines are cross-referenced in external resources such as EBI BioSamples IDs, ECACC catalog numbers and Cellosaurus accession IDs. The recording of cell line synonyms in hPSCreg also helps to find lines by their alternate names, sometimes well-known names for historical reasons, such as H9 or Shef-5.
hPSCreg uses ontology lookup services from EBI (ZOOMA) to allow the user to choose relevant ontology terms for clinical phenotypes and cell types. Additionally, genes and proteins are specified using Entrez Gene IDs (NCBI) or Ensembl Gene IDs (EMBL-EBI). Small-scale genetic information, such as HLA or genome variants, are collected using their respective standard formats, set out by the HLA Informatics Group (http://hla.alleles.org) and the Human Genome Variation Society (https://www.hgvs.org/), respectively.
Future developments in hPSCreg to increase its interoperability include ethics codification, implementation of stem cell community standards such as Minimum Information About a Cellular Assay for Regenerative Medicine (MIACARM) for stem cell line registry data exchange, and the linking of hPSCreg to other registries, such as the rare disease resource RD-Connect (https://rd-connect.eu/).

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

All cell lines registered in hPSCreg are assigned a unique standard name according to a stem cell community standard, which enables the unambiguous identification of a registered cell line across research outputs. hPSCreg itself as a stem cell registry is referenced at re3data.org (http://doi.org/10.17616/R36B9H), fairsharing.org (https://fairsharing.org/10.25504/FAIRsharing.7C0aVE) and Identifiers.org (MIR:00100898).

Outline naming conventions used

hPSCreg uses a standard stem cell nomenclature to name hPSC lines (https://dx.doi.org/10.1016%2Fj.stemcr.2017.12.002). The standard cell name itself contains information such as: 1) the generating institution; 2) the type of pluripotent stem cell (embryonic or induced); 3) primary cell line from a donor or a genetically modified cell line (termed subclone) from the same donor. As an example, BIHi004-C indicates an induced pluripotent stem cell line generated by the institution Berlin Institute of Health, the 3rd (C) cell cell line from the donor 4 at this institute. BIHi004-C-1 would indicate the first genetically modified line from the parental line BIHi004-C.

Outline the approach towards search keyword

Currently all text fields (except registered user data such as email and passwords) and metadata are searched using the user-supplied query terms.

Outline the approach for clear versioning

The data stored in hPSCreg is dynamic, as users are encouraged to complete and update their cell line data. Changes to the data are tracked in the "history" of a cell line, where the changes made, the user who made the change, and the time/date stamp are recorded by hPSCreg.

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

hPSCreg is working closely together with stem cell interest groups such as the International Stem Cell Banking Initiative (ISCBI) and the Global Alliance for iPSC Therapies (GAiT) to develop standards for stem cell data. hPSCreg is currently assessing the MIACARM (https://doi.org/10.5966/sctm.2015-0393) framework for data exchange with other stem cell registries.

2.2 MAKING DATA OPENLY ACCESSIBLE [FAIR DATA]

Specify which data will be made openly available? If some data is kept closed, provide rationale for doing so.

Most of the cell line data is made publicly available once the user who registered the cell line has filled out all mandatory fields (defined here: https://hpscreg.eu/docs/downloads/QuickStartGuideCellLineRegistration.pdf) and submits the cell line for validation. Briefly, cell line data that is made public includes:

- General information: the standard cell line name, alternative names, information about the institution that generated the line, publications associated with the cell line, cell line availability
- Donor information: genetic sex, clinical phenotypes, karyotype, details of the conditions of donor consent (ethics)
- Derivation: source cell type for iPSC reprogramming, reprogramming method, culture conditions
- Characterisation: marker expression for undifferentiated status and differentiation into three germ layers, morphology
- Genotyping: whether or not this data is available
- Genetic Modification (if performed): locus and modification method

Importantly, data that is not publicly released includes:

- User data: tracking history, identity of the user who registered the line. These data are stored for internal hPSCreg use only.
- Sensitive genetic data: HLA, STR, sequencing data. These data are kept under controlled access. An application to use these sensitive data must be submitted to the hPSCreg Data Access Committee, which decides how the data can be accessed, according to the provisions of the donor consent.

Specify how the data will be made available

Data about registered cell lines, research projects and clinical studies, which are deemed for public view, can be accessed on the hPSCreg website: https://hpscreg.eu .

Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

Only a web browser and access to the internet is required to access the publicly available data in hPSCreg.

Specify where the data and associated metadata, documentation and code are deposited

All data and associated metadata from the Registry are stored on virtual machines at the Charité - Universitätsmedizin Berlin. hPSCreg documentation can be viewed on the hPSCreg website (https://hpscreg.eu/about/documents-and-governance). The code that runs the Registry has not been made available in a public repository.

Specify how access will be provided in case there are any restrictions

Access to sensitive, personally identifying data such as genetic data is controlled by the hPSCreg Data Access Committee. Research groups interested in obtaining such data must make a formal application to the hPSCreg Data Access Committee. If access is granted, the requestor must agree to the terms of the Data Access Agreement. Data will be transfered to the requestor in a secure manner.

2.3 MAKING DATA INTEROPERABLE [FAIR DATA]

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

- Cell line identifiers: all cell lines registered in hPSCreg receive a unique, informative name that unambiguously identifies a cell line. hPSCreg also associates alternative names (such as those assigned by the generators of the cell line), as well as crossreferences such as BioSamples IDs, Cellosaurus accession numbers and ECACC catalog numbers, to facilitate the identification of the same line in different resources.

- Metadata vocabularies: hPSCreg uses ontology lookup service (EBI's ZOOMA) to allow users to choose controlled ontology terms for cell types, diseases and clinical phenotypes. Moreover, the standard stem cell line names have also been incorporated into Cell Line Ontology (CLO; https://doi.org/10.1186/2041-1480-5-37).

- Stem cell data-specific metadata: hPSCreg is trying out MIACARM (https://doi.org/10.5966/sctm.2015-0393) as a standard to exchange stem cell line data with other registries.

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

Presently, not all data in hPSCreg is codified or annotated with metadata. One of the next goals of hPSCreg is to codify the ethics section, possibly using guidelines from the Global Alliance for Genetic Health (https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/).

2.4 INCREASE DATA RE-USE (THROUGH CLARIFYING LICENSES) [FAIR DATA]

Specify how the data will be licenced to permit the widest reuse possible

The owners of the cell lines that are registered in hPSCreg retain ownership of the cell line data. Upon submitting the cell line data to hPSCreg, the user agrees on the behalf of the owners of the cell line data to make the cell line data public. The Registry does not place any additional restrictions on the use of the public cell line data. Sensitive genetic data, however, is subject to access through a Data Access Committee.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

The user who registers the cell line controls the public release of the cell line data by: 1) not submitting the data, in which case the data does not go public on hPSCreg; 2) setting a hold date and submitting the data for validation, in which case the cell line data is withheld from public view until the embargo has expired; 3) submitting the data, in which case the data goes public on hPSCreg. Once the data is public on hPSCreg, is it available for re-use.

Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why

All data that is publicly shown on hPSCreg is useable by third parties. hPSCreg aims to become a sustainable project to continue to provide a registry of qualified hPSC lines and clinical study information.
hPSCreg also aims to provide and document a stable API for data access in machine readable form.

Describe data quality assurance processes

Cell line quality: the user-supplied biological data is used by hPSCreg to assess the pluripotency of the cell lines following standards set by the stem cell community. We use the change history to track down and revert potential errors in the data.
Ethical provenance: conditions of donor consent are reviewed by hPSCreg Administration, based on the user-supplied data. The hPSCreg Ethics Advisor is consulted in cases where insufficient data has been provided by the user. Standard Operating Procedures for validation of the cell line quality and ethical provenance are on the hPSCreg website (https://hpscreg.eu/docs/downloads/hPSCreg_SOPs.pdf).
Clinical study data: following data entry of clinical trial data into the Registry by hPSCreg Administrators or external users, principal investigators of clinical trials are contacted by the hPSCreg Coordinator to confirm the existing data and are requested to provide missing data, if any.

Specify the length of time for which the data will remain re-usable

The data in hPSCreg will remain re-usable for the duration of the project funding, after which a sustainable alternative must be found to keep the server running in a secure way and foster hPSCreg's growth and adaption.

3. ALLOCATION OF RESOURCES

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

hPSCreg is involved in two EU-funded projects that can fund bio/informatic positions to further FAIRify hPSCreg data. It is estimated that 12 PMs per year will be needed to work on FAIRification. As a data project, the FAIRification tasks are integrated into the overall maintenance and development of hPSCreg.

Clearly identify responsibilities for data management in your project

The roles and responsibilites for data management lie as follows:
The hPSCreg Coordinator:

- oversees all operations of the Registry, including data management
- fosters collaboration with other registries for data exchange.

The hPSCreg Project Team:

- maintains and develops the Registry, including front and back-end development, regular data backups, system and software upgrades.
- implements measures to keep the data as secure as possible.
- implements measures to promote metadata capture and increased interoperability
- performs quality checks on the user-provided data, such as completion of all mandatory fields and validation of cell line data for certificates.

The hPSCreg Ethics Advisor:

- ensures the on-line data entry form for ethics and consent correctly captures sufficient information to evaluate the ethical provenance according to EU standards
- checks ethical provenance of cell lines in cases where documentation may be missing

The Charité - Universitätsmedizin Berlin IT staff:

- provide infrastructure to operate the host servers in a secure manner.
- provide infrastructure for regular backups.
- provide infrastructure for secure data transfer to third parties when required.

The registered hPSCreg users:

- follow the General Data Protection Regulation (GDPR) to protect the donor and cell line data. This may include anoymisation or pseudonymisation of the data.
- ensure that they have consent from the donor of the cell lines to make the anonymised or pseudonymised data public, for example, by registering and submitting the lines to hPSCreg for on-line publication or by publishing the data in a scientific journal
- ensure that donor consent covers the conditions of access to genetic data (no access, open access, controlled access)
- enter the data and metadata into the Registry using the on-line web form

- provide hPSCreg with truthful and correct scientific data about the cell lines and their ethical provenance.
- update their user data and cell line data if there is new information.

The hPSCreg Data Access Committee:

- evaluates external applications to sensitive genetic data.

The requestors who wish to gain access data that are under data access control:

- agree and adhere to provisions of the hPSCreg Data Access Agreement.

Describe costs and potential value of long term preservation

The mid-term (ca. 5 years) preservation of the hPSCreg data beyond the funding period for the hPSCreg project is subject to the infrastructure services of the IT Department at Charité - Universitätsmedizin Berlin, which currently hosts the servers that run the Registry. For a long-term solution staff is needed to update and adapt the systems with supported software versions to keep the server running in a secure manner. The amount of work needed in the future is difficult to predict, because it depends on the software “ecosystem”, but can be small when the system is designed well (as low as 1 PM every 3 years). It is unlikely that the project server persists for more than 10 years without proper maintenance and parts of the system are prone to fail in the first years without maintenance. hPSCreg has interactions with other systems and depends on user-entered data. These factors increase the level of required maintenance for functionality.

Long term preservation of the Registry data is of great value to:

- researchers who require certificates for qualified lines in their EU-funded projects
- funders and regulators who can use the Registry information to track research outputs of hPSC lines
- industry partners who are looking for suitable lines for further development towards cell-based assays or therapies

4. DATA SECURITY

Address data recovery as well as secure storage and transfer of sensitive data

Regular scheduled backups of the data and software running hPSCreg are saved on-server and off-server. Data in hPSCreg are stored only with necessary file system permissions and the server is secured from online and offline access.. No unauthenticated access is possible and only the actual HTTP(S) port is open to the public. No direct database access is possible from the outside. Data transfer is done via encrypted channels like SSH and HTTPS. The system is encapsulated in a virtual server that is not used for other systems.

5. ETHICAL ASPECTS

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

There should not be any legal or ethics issue regarding the public release of anonymised or pseudonymised donor of cell line data, as this is not personally identifying information. Sensitive genetic data, such as HLA, STR or sequencing data could in principle be used to re-identify the donors. The access to this data is controlled by the hPSCreg Data Access Committee.

6. OTHER

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

Not applicable.