



## RESPIRE Data Management Plan (DMP): Template (adapted from the University of Edinburgh)

Name:	Dr. Harshpreet Kaur
Modified by:	Tathagata Bhattacharjee
Project Title:	Formulating and testing a strategy for introducing Pulmonary Rehabilitation (PR) for COPD management in a rural Indian setting
Institute:	KEM Hospital Research Centre, Pune
Start Date:	01 Oct 2019
End Date:	31 Dec 2020
DMP version number and date:	V1, dated 23 October 2020

### **Responsibilities & Resources (applicable across the sections below)**

#### ***Who will be involved in the data management of this research?***

For efficient data management of PR study, different categories of resources will be involved, inclusive of human and other types.

The roles of human resources involved are:

1. Field Research Assistants (FRA) for on the field work
2. Field Research Supervisors (FRS) for supervisory work on and off the field
3. Counsellor for the psychosocial aspects for the PR activities.
4. Behaviour Scientist for developing and implementing the Behaviour Change Communication (BCC) component.
5. Physiotherapy Assistant for the physical training required for the Pulmonary Rehabilitation activities.
6. Data Entry Operators (DEO) for transferring the paper form data into electronic format
7. Project Manager (PM) for overall coordination and management of the study
8. Data Manager (DM) for electronic data management
9. The Principal Investigator (PI) for overall responsibility of data generation

The individuals assigned the specific roles described above are:

1. Babu Pawar in the role of FRA will approach the patients to invite them for study participation.



2. Nandedo Kale in the role of FRS will carry out the technical and field work and also help in note taking during the qualitative data collection and transcriptions of the same.
3. Counsellor (Shalan Mhetre)- responsible for the psychosocial aspects for the PR activities.
4. Behaviour Scientist (Dr. Harshpreet Kaur and Diksha Singh) responsible for answering all questions raised by the study participants, qualitative data collection and analysis for the needs assessment as well as for developing and completing the BCC component of the study.
5. Physiotherapy Assistant (to be decided) responsible for the physical training required for the Pulmonary Rehabilitation activities.
6. Data Entry Operators (to be decided) responsible for data entry of collected data. Please note that, although study budget has no provision for data entry operators KEMHRC will support for the same from their internal budget.
7. Aditi Apte as the project manager will be responsible for overall coordination of the study.
8. Sandeep Bhujbal as data manager responsible for overall management of study data. Please note that, although study budget has no provision for data manager KEMHRC has supported for the same from their internal budget.
9. Sanjay Juvekar will be responsible for project data generation, safety, storage and data use.

Other resources utilised for data management processes during this study include the following:

1. Computers (desktop and laptops) will be used for data entry work.
2. External hard drives for backup during the data entry work.
3. Backup servers for storage of data at Vadu site, KEMHRC and cloud services.

At the end of this project data will be submitted to Edinburgh DataShare (<https://datashare.is.ed.ac.uk/>) for sharing data in public domain and for long term preservation on DataVault (<https://www.ed.ac.uk/information-services/research-support/research-data-service/after/datavault>).

## 1. Data Capture

### ***What data will be generated or reused in this research?***

The following data will be generated in the PR study:

1. **Needs Assessment Questionnaire**- Quantitative and Qualitative; will be stored in csv format. This will include data obtained from Stakeholders (n=16) and Patients suffering from COPD (n=16). The interviews will be audio recorded and transcribed using transcription aids like otter.ai, the transcripts will be saved in Microsoft word. Till date, data from 4 stakeholders and 4 patients has been collected.

2. **Clinical assessment data**- Semi quantitative data: will be stored in CSV format and text format (scanned pdf files). This includes study participants medical history, data about their illnesses, overall health status and risk factors. This data will be collected from the COPD patients in study area. Approximately, data from 60 patients will be collected.
3. **Weekly progress data**- Semi-quantitative data; will be stored in CSV format. This includes the performance of the participant after every PR session. This data will be collected from the COPD patients in study area who avail the PR service. Approximately, data from 60 patients will be collected.

The above-mentioned data will be collected in hard copy forms and would be destroyed after minimum five years from the protocol defined end of study point. Fully anonymized (without any identifiers) soft copy data will be stored by KEMHRC as per its data sharing and archiving policy which is in line with the guidelines set by the national (Indian) data access and sharing policy.

***How much data will be generated?***

A total of approximately 152 records will be generated for each type of data stated in the section above. Overall size of the soft copy data will be approximately less than 2 Gigabytes.

## **2. Data Management**

***How will the data be documented to ensure it can be understood?***

Each type of data as stated in section above will be collected using well-structured questionnaires. The questionnaires had instructions with pre-question, the literal question and post-question information sections. These questionnaires (blank templates) both in soft (pdf) and hard copies will be stored at site along with the soft copy datasets and hard copy filled forms respectively for any future references.

All hard copy data with respect to this study and its documentation, inclusive of codebook wherever applicable will be kept for a minimum of five years from the protocol defined end of study point.

Data quality control will be done using Stata v15 tool and appropriate labels will be assigned wherever necessary for manual quality monitoring and checks. However, in the final data formats, which are csv and text, the labels would not be included.

It is planned to do proper Data Documentation Initiative conforming to international standard for describing questionnaires, statistical data files. This will add detailed and quality metadata for the datasets generated in this study. Metadata documentation is planned to be done for study datasets. The plan for this data documentation includes complete study documentation along with all the processes and standards incorporated

and adhered to along with the other data metrics as will be identified during the process. A detailed variable level metadata will be created for easy end-user understanding at any point of time.

### ***Where will the data be stored and backed-up?***

Data will be stored and backed up as per the below listed principles:

1. All data generated in relevant retrospection, joint studies and partnered projects under RESPIRE may have a cleaned and anonymized subset copy on the UoE data repository, named as DataShare. Access to such data on DataShare will be public.
2. A copy of all data that is uploaded on DataShare will be retained by KEMHRC on an “as-it-is” basis along with a master mapping record for identifiers. This is needed for regulatory purposes. The copy retained at KEMHRC will not be uploaded on any other public access data repository unless agreed by both the UoE and KEMHRC.
3. All data generated in relevant retrospection, joint studies and partnered projects under RESPIRE may be put in Edinburgh’s DataVault, for long term preservation, however the copies on DataVault must be anonymized with master mapping data for the identifiers in custody of KEMHRC. Dataset on DataVault must have controlled access with a definite lifetime assigned as per institution’s policy.
4. For all datasets pushed on to DataVault, a copy will be retained by KEMHRC with assigned lifetime as per institution’s policy (5-8 years for KEMHRC) along with the master mapping data for identifiers. The location of storage and related services will solely be the responsibility of KEMHRC.
5. All in-process data, i.e., active research data, that may need sharing with group members remotely may be put on UoE’s DataStore (<https://www.ed.ac.uk/information-services/research-support/research-data-service/during/data-storage>). These types of data sharing will be guided by the MoU and data sharing agreements of the collaborating institutions.
6. KEMHRC’s document server may also be used for all in-process, i.e., active research data that needs sharing whilst working collaboratively within office premise local network or VPN.
7. All data on either DataShare, DataStore or on DataVault, the ownership lies with KEMHRC with grant of custody given to UoE under terms and conditions of MoU.

Based on the above principles, data generated from the PR study is/will be stored as described here:

1. The data storages of KEMHRC includes the following and all data stored are catalogued using standard methods and are considered as “enclaved”, meaning that no direct access would be given. Probable users can search from the catalogue and raise a request for copy of the data.
2. KEMHRC data storage server is located in Pune office. This storage server is a well configured secured storage for all project data and catalogued and accessible over local network only. These are not publicly available resources and are accessible from within the network in office premises.

3. The KEMHRC data storage server is also configured to serve as a document server and all in-process, i.e. active research data, that needs sharing with group members can be used for access from within the local network or over VPN.
4. KEMHRC data storage server located at Vadu office. This is a temporary storage server for storing in-process data and does not store the final archival versions and accessible over local network only. These are not publicly available resources and are accessible from within the network in office premises.
5. A complete copy of raw data permanently archived in the above-mentioned KEMHRC data storages and catalogued for a minimum period of eight years in order to comply with the KEMHRC data policy, IT laws of India and funder/sponsor requirements.
6. In-process data, i.e. active research data, if needed, may be put on UoE's DataStore (<https://www.ed.ac.uk/information-services/research-support/research-data-service/during/data-storage>) in cases of distributed teams to share files anywhere and with anyone with study groups.

### **3. Integrity**

#### ***How will you quality assure your data?***

1. Quality check of the questionnaire data will be done by the behaviour scientist.
2. First level, behaviour scientist who is involved in data collection will check the collected data for completeness and logical checks.
3. Once data is collected and checked, it will be again reviewed by another behaviour scientist for the quality checks by using the pre-defined criteria.

### **4. Confidentiality**

#### ***How will you manage any ethical and Intellectual Property Rights issues?***

All Investigators and study site staff involved with this study conformed with the requirements of the General Data Protection Regulation (GDPR) 2018 with regard to the collection, storage, processing and disclosure of personal information and uphold the Act's core principles. Access to collated participant data is restricted to individuals from the research team, treating physicians of the participants, representatives of the sponsor(s) and representatives of regulatory authorities.

Computers will be used to collate the data and have limited access measures via usernames and passwords.

All identifying information that will be collected about the participant (such as name, age, sex, address, contact information) during the course of the research is kept confidential

and secured. Published results will not contain any personal data that could allow identification of individual participants.

The data of each study participant will be identified with the help of a unique identifier and it will be completely anonymized and scrambled before sharing. The details of the unique identifier will be held with the research team. There will be no such information in the shared data which will disclose the identity of the study participant. Standard and recommended security measures and confidentiality with data sharing agreements will be in place with access control at every stage and audit trails maintained for all access and changes in data.

## 5. Retention and Preservation

### ***Which data do you plan to keep and for how long?***

Data will be retained and preserved as per the principles stated in section above in Data Management.

All hard copy data (filled forms), which includes identifiable information and related documentation will be preserved at KEMHRC Vadu for up to a period of five years from the protocol defined end of study point. After the elapse of five years, hard copy data will be destroyed as per KEMHRC guidelines and/or specific contract clause with the sponsor(s), if any or under prevailing law of the land (India).

Soft copy of the raw data will be uploaded on secured KEMHRC data storages with limited access to KEMHRC data administrators only. Data on KEMHRC data storages are catalogued. Data is “enclaved” in the storages, meaning it is findable through the catalogues but no direct access is given. Data is categorized and some categories of data, for example the identifiers, which are for internal reference only will not be made accessible to non-KEMHRC entities. The categories of data meant for public access either open or controlled will not be on these storages.

Any access needed is to be directed through the data administrator after due approvals. As per KEMHRC policy, this soft copy of data will be retained on the storage server(s) for a minimum period of eight years with no upper limit defined.

An anonymised copy of the study data will be backed up on the UoE’s DataVault (<https://www.ed.ac.uk/information-services/research-support/research-data-service/after/datavault>) for long term preservation. The preservation details are articulated under the next heading.

### ***How will the data be preserved?***

Based on the principles listed in section 2, data generated for the PR study is/will be preserved at the end of study as described here:

1. Soft copies of all data collected in PR study is/will be anonymised with identifier mapping master.



2. Soft copies of all data is/will be preserved by KEMHRC along with the mapping master which will be retained as per data policy of KEMHRC (The KEMHRC data policy is not made available as public accessible resource as on date; however, it is sharable with collaborators on approvals from the trust members).

3. Data will be preserved on University of Edinburgh's DataVault (<https://www.ed.ac.uk/information-services/research-support/research-data-service/after/datavault>) for a longer period as defined by the University's data policy.

4. KEMHRC will preserve data on its data backup servers located in KEMHRC Pune office and also on commercially purchased data archival cloud space (<https://aws.amazon.com/glacier/>).

5. All soft copies of data including identifiable information and related documentation will be preserved on KEMHRC storages and anonymised copies on UoE's DataVault (<https://www.ed.ac.uk/information-services/research-support/research-data-service/after/datavault>).

A complete copy of the anonymised data validating the results is/will be preserved for long term in the above-mentioned data storages and catalogued for a minimum period of eight years in order to comply with the KEMHRC data policy, IT laws of India and funder/sponsor requirements.

## 6. Sharing and Publication

### ***Which data will be shared and how?***

Data sharing principles encourage ethical commitments of data generated from the public and must benefit the public by sharing for open access research opportunities. KEMHRC holds this principle to its core and provides data from its studies and projects for sharing after due processes of cleaning, anonymisation and masking confidential information wherever applicable.

For PR study, KEMHRC would be submitting anonymised data for public access on University of Edinburgh's DataShare (<https://datashare.is.ed.ac.uk/>). Data on DataShare must follow the principles of findability, accessibility, interoperability, and reusability (FAIR) and the submitted dataset must have a Digital Object Identifier (DOI) assigned.

As per KEMHRC policy, data on KEMHRC server will be stored for a minimum eight years with no upper limit defined. A similar copy of the data would be retained by KEMHRC for adherence to local IT laws. Any derived or calculated or other form of data can also be shared on DataShare.

### ***Are any restrictions on data sharing required?***

There are a few restrictions and procedures for compliance to KEMHRC data policy and local IT laws:

- (1) Identities of study participants cannot be shared or stored on servers outside the boundaries of India
- (2) Only anonymised data can be shared on public domain. The degree to which anonymisation is done must be clearly understood and documented.
- (3) A copy of all data stored on servers outside India must have a copy within Indian territory and must be made available to any law enforcing or regulatory agency on demand
- (4) The law enforcement and regulatory authorities will have full access to the data as per the rules and regulations.

Not all of the above is law yet but compliance is solicited. It is expected that any researcher using this dataset for any type of publication or conference paper must cite this dataset by referencing the DOI.

====end of the document====