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

Name:	Dr Ankita Shrivastava
Modified by	Tathagata Bhattacharjee
Project Title:	To assess feasibility of method to estimates the costs of care, quality of life and wider societal burden (mortality, morbidity and lost productivity with economic impact on family) due to COPD
Institute:	KEM Hospital Research Centre, Pune, India
Start Date:	01 Jan 2020
End Date:	31 March 2021
DMP version number and date:	V1, dated 02 October 2020

Responsibilities & Resource (applicable across the section below)

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

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

The roles of human resources will be involved are:

1. Field Research Assistants (FRA) for on the field work
2. Project Manager (PM) for overall coordination and management of the study
3. The Principal Investigator (PI) for overall responsibility of data generation

The individuals assigned the specific roles described above are:

1. Vaibhav Raut, Seeta Jadhav, Swati Shivale and Archana Gundal are the FRAs who will be doing the data collection.
2. Ankita Shrivastava will be the project manager with responsibilities for overall coordination of the study.
3. Sanjay Juvekar will be responsible for project data generation, safety, storage and data use.

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

1. Computers (desktop, laptops, tablets) used for data collection and 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?

Data generated in this study are summarised as follows:

1. **Questionnaire:** Quantitative; stored in csv format. This will include study participants' demographic information, cost of illness data and quality of life assessment. This data will be collected from the COPD participants which were identified in the 4CCORD study. Data will have information about how and how much they spend (directly or indirectly) on their health condition and its impact on their routine life.

2. **Qualitative data**: In-depth interviews will be done with few COPD patients and healthcare providers using an interview guide tool. Very detailed information from the COPD patients will be collected on their health care expenditure including the impact of the same on their routine life. From health care providers, detailed information about financial implications of COPD on the health system and its impact will be collected. Interviews will be initially audio recorded and then converted into transcripts after matching with the notes. Transcripts of these interviews will be stored in PDF format.

The above-mentioned data will be collected in hard copy forms and will be destroyed after a 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 120 records were generated for cost of illness and quality of life data stated in the section above. Along with this, approximately transcripts of 16 in-depth interviews will also be stored. 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?

Cost of illness and quality of life assessment data as stated in section above will be collected using well-structured questionnaires. The questionnaires will have 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. Qualitative data be collected through interview guide.

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 will not be included.

It is planned to do proper Data Documentation Initiative conforming to international standard for describing surveys, questionnaires, statistical data files, and social sciences study-level information. 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/preserved/shared 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 is 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 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 must 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 for the said 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 with 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 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?

Quality checks of the questionnaire data is done at three levels.

1. FRAs involved in data collection will check the collected data for completeness, accuracy and logical checks. Project Manager will perform random quality checks for a few questionnaires.

2. For qualitative data, transcription will be done by the individual who has collected the data, and 10% of transcripts will be compared with audio recording for quality check by Project Manager.

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

All identifying information that is 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.

Data collected through in-depth interviews will not include personal identifiable data and will be anonymised at the time of transcription. We will use pseudonyms or replacements or codes throughout

the study. The unedited versions of data will be kept with selected researchers in the research team. There will be an anonymisation log of all replacement, removals, and codes and will be stored separately from the data files. Audio recordings collected through encrypted device, will be transcribed immediately after the interview by the study team.

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 is 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 is 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 for long term preservation:

https://www.ed.ac.uk/information-services/research-support/research-data-service/after/datavault

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 said study will be preserved at the end of study as described here:

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

2. Soft copies of all data 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 will be preserved for long term in the above-mentioned data storages and catalogued for a minimum period of eight years 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 this 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====

This research was funded by the National Institute for Health Research (NIHR) Global Health Research Unit on Respiratory Health (RESPIRE) using UK aid from the UK Government to support global health research.

The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the UK Department of Health and Social Care.