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

Name:	Dr. Osman M Yusuf
Project Title:	Estimating Chronic Respiratory Disease (Asthma and COPD) burden in adults in Asian low and middle-income countries [4CCORD study - 4 Country ChrOnic Respiratory Disease study] PILOT PHASE, Islamabad
Institute:	The Asthma & Allergy Institute Pakistan
Start Date:	01 Sep, 2018
End Date:	31 Aug, 2019
DMP version number and date:	V-1.0 14 October 2020

Responsibilities & Resources (applicable across the sections below)

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

Data will be maintained, after vetting by the Principal Investigator and by the Data Management Team, under Dr. Aimal Rextin for entering, maintenance, and security.

1. Data Capture

What data will be generated or reused in this research?

This overarching project collected health data from 4 different countries. The main research team was based in India and was led by Prof Sandeep Salvi. While AAIP was responsible to collect data from Pakistan and pass it to the research team in India for detailed analysis. The data was emailed to Prof Sandeep's team as agreed in the protocol approved before the start of the project.

In this project, we collected clinical data of two groups of patients. We initially recruited 100 participants and collected some initial data. This data was then passed to the team led by Sandeep Salvi in India, who requested AAIP to collect further details from 43 of these 100 participants. Hence the 43 participants from whom further data was collected was a subset of the overall 100 participants that were initially recruited by AAIP. We will describe them one by one below:

Group of 100 persons

For this group of persons, we collected the following information:

- Clinical histories of study participants stored in excel files including sensitive information like gender and age.
- The results of their breathing tests (Spirometry readings) as excel graphs converted to PDFs.

Subgroup of 43 patients

Out of these 100 people, 43 were recommended by the researchers (Prof Sandeep Salvi) for further investigation, and the following additional data was collected.

1. X-ray reports as image files.
2. X-ray images
3. Clinical notes of specialist doctors when they examined these group of patients. These will be image files of their handwritten notes.

All highly sensitive information in clinical histories like name and phone number, address etc have been replaced with numeric codes to avoid any risk to privacy when sharing it outside AAIP. We were assigned a series code from 5001 to 5100 by the study coordinators in India. So, each person was awarded a unique ID between 5001 to 5100. (No 5 was awarded as the country code for Pakistan). Sensitive information like name, phone number etc is only known to Dr. Osman Yusuf.

The identifier data has been securely stored in a password protected folder in the laptop of the PI, with the password known only to him. The hard copies are under lock and key is under his personal supervision.

How much data will be generated?

1. 100 Questionnaires filled in and entered in Excel. – Total Size= 101.6 KB
2. Spirometry readings - 100 pdf files - Total Size of 100 files = 7 MB
3. Data of 43 patients selected for further study:
    i. X ray scan images - 272 KB each file (jpg file)
    ii. X ray report images - 688 KB each file (jpg file)
    iii. Clinical notes images - 312 KB each file (jpg file)

Total memory requirement = $43 \times 1272\ KB + 101.6\ KB + 7000 KB \approx 10\ MB$

2. Data Management

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

We will generate the following:

1. A technical report will be prepared for the secondary users to understand the collection procedure and processing of the generated data with a bibliographical citation for users to cite in future publications. This report will

also explain the operational use of the variables.

2. A readme.txt file to explain how the datasets relate or are linked to each other. All documentation and metadata will be generated in a document (.txt) and published by Edinburgh DataShare with the files described above.

The metadata will be submitted to the archive of University of Edinburgh, i.e. DataShare in a format relevant to their metadata requirements/ standard.

Where will the data be stored and backed-up?

Various file formats like excel, jpeg, and pdf as elaborated in Section 1.

The data files for the project will be structured with a root folder in which the common files for all 100 persons will be stored. In the directory there will be a sub directory in which the additional information for the 43 suspected patients will be stored.

These files are stored on the local hard drive of a computer at AAIP. These files are also backed up on a shockproof external local hard disk dedicated to the project and is password protected to ensure its security. The password of these files will be shared with a very limited number of concerned personnel working for the project, but currently only with the PI.

Once the project is completed all sensitive information except the age and gender of the participants will be removed and stored on DataVault to ensure that the data is accessible only to authorized scientific research groups. Age and Gender were retained as they are essential to plan future surveys and interventions. If these variables were deleted than a researcher will not be able to infer how age/gender influence the values of spirometry.

The metadata and the data in aggregated form will be stored on DataShare to ensure maximum possible availability.

3. Integrity

How will you quality assure your data?

Overall, the quality assured by the PI who oversaw all aspects of the data collection, data entry and analysis. More specifically, quality assurance at AAIP for the 4CCORD Project was achieved through the following steps.

Step 1: Study Design

The sampling technique was designed by the Project Manager, who holds an MPH and PhD in Epidemiology. The project manager was based in Allergy and Asthma Institute Pakistan, Islamabad. This was cross checked by the Principal Investigator and verified by a statistician.

Step 2: Data Collection

Quality assurance during the data collection was achieved in the following manner.

1. Every data collection team comprised of 2 field researchers (one male and one female) who cross checked and verified every piece of information collected at the time of data collection.
2. Spirometry blowing techniques were verified by the spirometer itself, giving information about incorrect blows.
3. All spirometric readings were then verified by Prof Sandeep Salvi, Consultant to this project and all doubtful readings were repeated on his advice.
4. An independent supervisor visited the data collection sites at a later date to verify the visit, data collection, and randomly verify the questionnaires. A consultant pulmonologist from the Pakistan Institute of Medical Sciences, Islamabad was hired from this purpose and his work was coordinated by AAIP. The consulted signed an agreement at the time of his joining the project that contained clauses of non-disclosure. He unfortunately passed before the project finished.
5. Dr Shahida Ashraf (Co-PI) independently verified the visits, data collected and asked about complaints (if any) from the persons sampled.

Step 3: Data Entry

The data was entered in duplicate in Excel by the Project Manager and the Data Entry person, and correlated for any errors, missing data, duplication etc. The data was again checked by Dr Sandeep, who returned the PDF files with handwritten notes if he found them problematic.

Step 4: Clinical Assessment

Dr Sandeep Salvi assessed each patient for the clinical features and advised further investigations (if required). The consultant Pulmonologist (Dr Mudassar Shafiq from PIMS) verified each suspected patient clinically. Dr Shafiq has a healthcare practice in Rawalpindi, Pakistan. The PI checked on the clinical features of 21% randomly selected patients (9 out of 43).

4. Confidentiality

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

The study coordinators (Prof Sanjay Juvekar and his team in India) obtained study ethical approval from ACCORD. Furthermore, local Pakistani clearance was obtained from the International Research Force.

All human participants in this study were informed about the data being collected, the purpose of the study and that it might be shared with other researchers after anonymization. Highly sensitive information like name and address were removed at an early stage and each participant was given a unique identifier. All data concerning a participant will also contain that unique ID known to members of the team. Hence, the

data will not contain any personally identifying information, however, we will know that when two data entries concern the same person.

The data will still contain some sensitive variables like age and gender however it will be shared in aggregated form. In non-aggregated form, the data will be available only on DataVault to ensure that that the data is not misused.

The data generated in this project will be entirely co-owned by AAIP and the University of Edinburgh. However, it will be available to be used for research purposes.

5. Retention and Preservation

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

The dataset with age and gender will be maintained at AAIP in a secure computer indefinitely for possible future scientific utilization. The data will also be preserved on DataVault data repository for 10 years and its aggregated version will be preserved indefinitely on DataShare.

How will the data be preserved?

Since, the data contains personal information like age and gender so it will be available through DataVault due to its privacy and security features. However, the aggregated data will be available without any restrictions through DataShare.

6. Sharing and Publication

Which data will be shared and how?

The data of the results of the study have been shared with the data management team in India through email. They will be responsible to upload and/or share any data which has to be shared with the UoE or any external partner.
The AAIP (4CCORD's Pakistan centre) is only responsible for publishing its own data, after the 4-country data has been published by the 4CCORD central team led by Dr. Sanjay Juvekar. As discussed, earlier AAIP will share its data with other research teams through DataShare in aggregated form and DataVault in non-aggregated form. This will be published in relevant peer-reviewed journals and shared in the form of a country report with relevant health authorities, administrators and other stakeholders.
It may be presented online at a suitable conference, if funds are available.

Are any restrictions on data sharing required?

There will be no restrictions on the aggregated data and the metadata that is available on DataShare.

The access to the full data available on Datavault will be provided only after a formal request and evaluation process. The purpose of the evaluation would be to ensure that the data is shared only with scientific community and unauthorized persons will not be able to access and use the data.