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

Name:	Dr. Osman Mohammad Yusuf
Project Title:	To assess the feasibility of using the eDPSEEA model in seasonal pollen induced asthma in Islamabad
Institute:	The Asthma and Allergy Institute Pakistan
Start Date:	1 Sep 2019
End Date:	28 Feb 2021
DMP version number and date:	V-2.0 November 2020

Responsibilities & Resources (applicable across the sections below)

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

All aspects of data management including its integrity, maintenance, and security is the responsibility of Dr. Aimal Rextin.

1. Data Capture

What data will be generated or reused in this research?

Broadly, there are two different sources of data. The first data source is clinical data which starts as paper records and is being entered in electronic format by personnel dedicated to data entry. The government hospital that is part of this study has agreed not to record any personally identifiable information on paper. The second data source is nonclinical data like pollen counts, weather data, air quality data etc.

Most of this data is numerical in nature and the rest is short texts. Hence, .csv would be the most suitable format as it will not restrict the data to any platform. There will also be image and video data that will be saved in general formats (.jpeg, bmp, mp4 etc.) so they are universally accessible. The personally identifiable information in the clinical data will be replaced with numeric codes and aggregated to avoid any privacy risk when sharing it outside AAIP.

Specific sources from which data is being collected are the following:

1. Clinical Data:

a. Data from the cohort of 40 asthma patients
    i. Basic information such as age, gender and background medical history
    ii. Date wise record their asthma situation.
    iii. All data will be encoded in .csv files for electronic storage.

Details

AAIP has recruited a cohort of 40 asthmatic patient who are being monitored throughout the project for detailed feedback. All participants of this cohort have given their informed consent to become part of this scientific study and the sharing of their anonymized data for possible future research studies. Data is being collected and maintained by Dr Ashraf. His main responsibilities include regular receipt of questionnaires, using telephone calls, and personal visits of the patient for detailed feedback when required.

b. Retrospective Clinical Data of 100-150 patients from The Allergy & Asthma Institute
    i. Date wise record of their medical history at AAIP.
    ii. All data will be encoded in .csv files for electronic storage.

Description

This will be retrospective data of 100-150 patients from Allergy and Asthma Institute, Pakistan (AAIP). These will be randomly selected out of 400+ patients who have undergone allergy testing over the last 20 years and presented with asthma and/or nasal allergies. This data will be collected from scanned paper records transcribed from 10 randomly selected years (not consecutive) by Dr. Ashraf and the Medical Officer and Data Transcriptionist based at AAIP. We will use codes to keep the identity of the respondents anonymous.

3. Data of asthmatic patients at Pakistan Institute of Medical Science, Islamabad (PIMS).
    i. Date and number of patients given nebulizers.
    ii. This stored is being stored as .csv files.
    iii. No medical or personal information of the patients is being recorded.

Description

Medical staff members are recording the number of patients given who were given a nebulizer every day. Four staff members were employed to count the administration of nebulisers 24 hours, but this has been reduced to two staff members due to the COVID situation.

4. Data of pollens and spores etc. collected from the project's Burkard samplers
Three Burkard pollen counters have been installed in Islamabad to sample pollen grains and other airborne particles, including fungal spores. We will soon start analysing the data microscopically and different types of pollen and spores etc. will be counted. These counts will be recorded electronically in .csv files. Images and videos of this microscopic analysis will be also be recorded to assure the quality of these counts.

5. Micrographs and Scanning Microscopy Video Data

i. 60 photomicrographs of each slide from each sampler, measured at 5 fields per vertical traverse at alternate hours (5 fields vertically, after every 2 hours = 60 fields to be photographed) per 24 hours.
ii. A video of scanning microscopy for every slide. This high-resolution video will be at least a few minutes long.

Description

A slide for each 24 hours duration starting at 4:00 pm each day till 3:59 pm the next day (time may be changed according to weather conditions or other reasons). This slide will be observed under a microscope and images and videos of this observation will be recorded and stored. These images and videos will be used to independently verify our pollen counts to assure the integrity of the pollen data, as well as to serve as a database of images for pollen identification.

6. Pollen count data collected from the Pakistan Meteorological Department

Historical and current pollen count data previously collected at Pakistan Meteorological Department and meteorological data such as humidity, wind, temperature, and precipitation will be used.

7. Weather data, Air Quality data and pollution data from various sources.
These will be collected from publicly or commercially available resources for example weather data for Islamabad (purchased from MeteoBlue), publicly available WHO regional data to estimate viral circulation, publicly available daily PM2.5 readings of Islamabad.

How much data will be generated?

Non-Image Data

We are unable to exactly calculate the volume of data as the number of patients and the number of fields collected for each patient have not been determined yet. However, we estimate that we will need less than 5 GB of memory storage.

Image and Video Data

We estimate that each micrograph will take approximately 10 MB. Since we have three sampling sites, that will be working for a year, 5 images per slide and the images will be recorded at alternate hours so we the total memory requirement for the micrographs comes out to be:

$$5 \times 12 \times 3 \times 365 \times 10MB = 657\ GB$$

Similarly, we estimate that each 2 min high resolution video will occupy 200 MB memory. We will make approx. 1 -2 min video per slide for 365 days, so the memory requirement comes out to be:

$$365 \times 3 \times 24 \times 200MB = 5.256\ TB \approx 5.5\ TB$$

Total Memory Requirement

$$5.5TB + 657\ GB + 5GB = 6.16\ TB$$

2. Data Management

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

Some metadata such as questionnaires, technical notes, variables dictionary, etc. will be necessary. 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 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 into the DataShare repository with the files described in the above section.

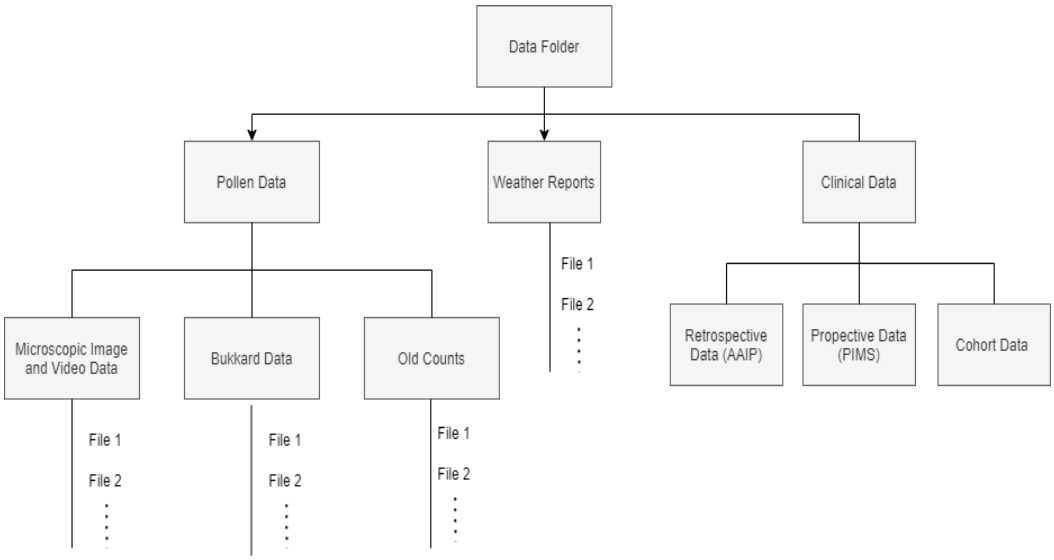
We will also include a note in each data file that indicates the location of stored metadata. The metadata will be submitted to the archive of University of Edinburgh, i.e. DataShare in a format relevant to their metadata requirements/ standards. DataShare was selected for its lower cost and our minimal requirements regarding protecting the data due to aggregation.

Where will the data be stored and backed-up?

The data files for the project will be structured according to the type of data, i.e. quantitative (pollen count and weather reports etc.) and qualitative (clinical data). The folders will be structured and named in the following manner:

1. The data in pollen count file will be organized in two subfolders one for newly collected data and the other for old counts. These count records will be stored in separate files named according to week and date of observation.

2. A folder called Multimedia will contain subfolders called micrographs and video that will contain properly labelled image and video files.
3. Weather report folder will also contain sub-folders named in accordance with week.
4. They will be organized in a single folder that contains the information about patients from different sources. As discussed in Section 1, the clinical datasets will consist of 1) cohort data of 40 patients 2) prospective data from Pakistan Institute of Medical Sciences (PIMS) and 3) retrospective data of 100-150 patients from AAIP.



Folders will be maintained according to date, authors name, status of the file and document name.

Pollen samples are being collected from multiple locations in Islamabad. These samples will then be examined microscopically at Quaid-i-Azam University (QAU) under the supervision of Dr. Mushtaq Ahmed. Dr. Mushtaq on behalf of QAU and Dr Osman on behalf of AAIP have already signed a subcontract in this regard that include all relevant RESPIRE conditions regarding quality control, data ownership, and confidentiality etc. This data will be recorded at QAU and stored in a password protected computer where all research data will be stored in encrypted form. The password will be available only to Dr Mushtaq Ahmed and Dr Osman. Moreover, the drive containing the datafiles will be encrypted using `VeraCrypt`. The data will be shared with AAIP using an encrypted flash drive on a weekly basis, the flash drive will also be encrypted using `VeraCrypt`. In case of an unlikely event of this USB drive getting lost or stolen, the data will still be safe on the computer at QAU. The data will then be replicated at AAIP on a hard drive on a password protected computer with the data being kept in encrypted form. These files will also be backed up on a shockproof external local hard disk dedicated to the project and

will be password protected to ensure its security. The pollen data files and its passwords will be shared with a very limited number of concerned personnel working for the project, but currently we plan they will only be known to the PI, Dr Mushtaq Ahmed, and Dr. Aimal Rextin.

3. Integrity

How will you quality assure your data?

The following type of data will be collected during this project:

1. Clinical Data: As discussed above, all health information is being transcribed from paper records by dedicated personnel at AAIP. The data management team will select a random sample of paper records and match it with the electronic data to assess its integrity. This random sampling will be done on regular intervals (to be decided later) to assess the quality of the batch, a batch may be returned if the number of inconsistent records fall fellow a particular threshold that will be decided later by the PIs.
2. Pollen Data: In this regard, a team has been engaged to maintain the pollen counter, collect tapes, make slides, make micrographs and do the pollen counting. These pollen counts will be matched with random sample of slides to ensure accurate and meticulous counting. There will be two levels of oversight and quality assurance, first by Dr Mushtaq/Prof Zafar (in QAU) and the second by Dr Osman Yusuf/Prof SM Hasnain.

4. Confidentiality

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

We have obtained the approval of this study from ACCORD, furthermore, local Pakistani clearance has been obtained from IRF.

All human participants in this study are being informed about the data being collected, the purpose of the study and that it might be shared with other researchers after anonymization. Their data is only recorded and preserved if they give their informed consent. We give a unique identifier to each human participant, and so any data that is shared outside the core research team does not contain any sensitive personal data.

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

5. Retention and Preservation

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

All data mentioned above will be stored indefinitely after anonymization, i.e. we will destroy any personally identifiable data in the research data.

How will the data be preserved?

Since, the shared data will be anonymised, and the only personal information will be gender and age. Hence, we will use DataShare to share the data after aggregating it based on age and gender after the project ends.

6. Sharing and Publication

Which data will be shared and how?

Since, the weather data is collected from a meteorological service it cannot be further shared according to their conditions and policies. Data will be archived for long term storage at DataShare repository of the University of Edinburgh, which provides permanent dataset identifier (DOI) for deposited data. This data will contain pollen counts and clinical data aggregated based on age and gender.

All data will be encrypted and archived using an open source disk encryption software VeraCrypt and an archiver 7-zip before transferring electronically to the main repository. For active storage, all the data will be saved in password enabled files which will be accessible to the concerned researchers and Principal Investigators only.

Are any restrictions on data sharing required?

During the term of project, the data will only be accessible to the data team and the PI (Dr Osman Yusuf). For later research purposes, it will be shared according to the term of usage prescribed by the PI.