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

Name:	Mark Sun and Senjuti Saha
Project Title:	*Construction of a Computational Framework to Automatically Interpret Chest X-rays and Diagnose Pneumonia*
Institute:	Child Health Research Foundation
Start Date:	*01 July 2018*
End Date:	*30 December 2020*
DMP version number and date:	V0.1 December 06 2020

Responsibilities & Resources (applicable across the sections below)

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

Interpretation of paediatric chest X-ray images for the presence or absence of pneumonia will be performed by three clinicians/radiologists (label generation): Nawshad Ahmed, Mahamuda Apa, and Fatema Doza

Computational modelling (generation of the model and associated weights): Mark Sun

Analysis (results generation): Mark Sun and Senjuti Saha

1. Data Capture

What data will be generated or reused in this research?

The study objectives are 1) to create a high confidence set of labels (True / False) indicating if pneumonia is present or absent in paediatric chest X-ray images and 2) to use the high confidence labels to train a deep learning model to diagnose pneumonia directly from paediatric chest X-ray images. The images are derived from a WHO-supported surveillance study that has been on-going since 2013 and are managed under protocols governing the WHO surveillance study. The paediatric chest X-ray images are interpreted by at least two trained clinicians / radiologists for the presence of 1) primary end-point pneumonia (PEP), 2) other lung infiltrates, and/or 3) pleural fluid in either the left or right lung, for a total of six possible binary outcomes, which will henceforth be called “labels”. The resulting labels from the multiple readers are saved in an anonymized comma-separated (csv) file (the label file) and backed up to the secured CHRF server. To further ensure data persistence, the labels are additionally stored on 1) Google Drive whose at rest encryption uses 128-bit AES keys and whose in-flight encryption uses 256-bit SSL/TLS and 2) Microsoft OneDrive whose at rest encryption uses a 256-bit AES key and whose in-flight

encryption uses 2048 bit SSL/TLS. Statistical analyses to assess model and reader performance are conducted in R by importing the anonymized csv label file and the model predictions. Computational models to diagnose paediatric pneumonia directly from the X-ray images are developed in python. After a model is trained, the model architecture and weights are saved in files following the hdf5 file format. The model inputs are the X-ray images and the outputs are the predict labels. The labels and code will be provided in a plain text format for other groups to freely use. The model weights will be available only for non-commercial purposes.

How much data will be generated?

The label csv files will be less than 10MB in size. The model weights will likely be 500MB in size. Thus, the total amount of generated data will be 0 – 50GB.

2. Data Management

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

The data, code, models, and model weights will be documented via metadata stored in XML following the Data Documentation Initiative (v2.3) specification.

Where will the data be stored and backed-up?

The labels will be stored on a secured CHRF server and regularly backed up to Google Drive whose at rest encryption uses 128-bit AES keys and in flight encryption uses 256-bit SSL/TLS and to Microsoft OneDrive whose at rest encryption uses a 256-bit AES key and whose in-flight encryption uses 2048 bit SSL/TLS. The code, model, and model weights will be stored on the Amazon Web Services elastic file system service and accessible only from a single computer during model development.

3. Integrity

How will you quality assure your data?

To ensure that the radiological findings are robust, all chest X-ray images are interpreted by two readers for the presence of paediatric pneumonia, as defined by the WHO primary endpoint pneumonia definition. X-ray images with discordant findings are re-evaluated by a third reader. The kappa statistic between readers will be reported, to verify reader agreement.

Computational models that will diagnose pneumonia directly from the X-ray images will be trained on random subsets of the data (5-fold cross validation) to ensure model stability and data integrity.

4. Confidentiality

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

All the X-ray images were attained from an on-going WHO-funded Invasive Bacterial Vaccine Preventable Diseases (IB-VPD) Surveillance Platform in Bangladesh, where the X-rays were performed on the advice of the treating physician. Ethical and Intellectual Property rights follow guidance outlined by the IB-VPD study.

All labels arising from the interpretation of the chest X-ray images (i.e. the associated image labels), were performed by clinicians and radiologists recruited by the study authors and the data is anonymized before analysis. Consequently, no ethical or intellectual property (IP) rights issues will arise. All study protocols were approved by the Ethical Review Committees of the Bangladesh Institute of Child Health, Bangladesh, and all IP is held by Child Health Research Foundation, Bangladesh.

5. Retention and Preservation

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

The intent is to have the chest X-ray image labels, R / python code, and model architecture be perpetually available to the research community. Availability of the chest X-ray image labels will enable others to verify the data quality and may be retrieved from github and the Edinburgh DataShare. Availability of the R / python code and model architecture will enable others to verify the results and enable others to extend the findings to other yet unknown applications. The code and model architecture will be available on github and the Edinburgh DataShare. The model weights will be perpetually available to the research community for non-commercial use. The model weights will be stored on Microsoft OneDrive and Google Drive. These cloud-based storage solutions ensure that the model weights will be accessible to CHRF in the event of a multiple regional catastrophic events. All temporary files (e.g. model checkpointing files) generated during the project will not be kept as their creation is to enhance computational efficiency.

How will the data be preserved?

Upon completion of the study, the image labels, code, and model architecture will be publicly available on GitHub and on Edinburgh DataShare. The model weights will be stored on Microsoft OneDrive and Google Drive to ensure accessibility to CHRF, even in the event of multiple regional disasters. The model weights will not be available on a public data repository, reducing the legal costs required to enforce a license associate with the model weights. The research community may readily access the weights by requesting access, as the model weights are intended for non-commercial use.

6. Sharing and Publication

Which data will be shared and how?

Upon completion of the study, the image labels, code, and model architecture will be publicly available on GitHub. The image labels will also be available on Edinburgh DataShare. The model weights will be stored on Microsoft OneDrive and Google Drive.

Are any restrictions on data sharing required?

The image labels, code, and model architecture will be publicly available. The model weights will be provided to individuals or organizations only once written consent is attained from Child Health Research Foundation, Bangladesh, to ensure non-commercial use of the model.