# Data Documentation Template

## 1 Study Description

### 1.1 Identification

#### 1.1.1 Title

Contains the full authoritative title of the data collection. A full title should indicate the geographic scope of the data collection as well as the time period covered.

Vukuzazi Machine Learning Chest X-ray project among individuals aged 15 years and above in rural KwaZulu Natal, South Africa, 2018.

### 1.2 Overview

#### 1.2.1 Abstract

An unformatted summary describing the purpose, nature, and scope of the data collection, special characteristics of its contents, major subject areas covered, and what questions the PIs attempted to answer when they conducted the study. A listing of major variables in the study is important here.

The datasets provided here will be used for devising a machine learning tool to improve upon CAD4TB's triaging of chest x-rays.

#### 1.2.2 Kind of Data

The type of data included in the file: survey data, census/enumeration data, aggregate data, clinical data, event/transaction data, program source code, machine-readable text, administrative records data, experimental data, psychological test, textual data, coded textual, coded documents, time budget diaries, observation data/ratings, process-produced data, etc. No description just a single phrase, e.g. Genetic sequences

Survey data

#### 1.2.3 Unit of Analysis

Basic unit(s) of analysis or observation that the study describes:

Each chest x-ray image or images are for a Vukuzazi study participant who had chest x-ray done.

The clinical data, each record is for a single participant who participated in Vukuzazi and had chest x-ray done. Some individuals had more than one chest x-ray images and therefore have more than one record.

### 1.3 Scope

#### 1.3.1 Topics Classification

MeSH subject headings

Chest x-ray; machine learning; tuberculosis; population health screening>

#### 1.3.2 Keywords

Keywords summarize the content or subject matter of the survey. As topic classifications, these are used to facilitate referencing and searches in electronic survey catalogues.

Tuberculosis;

### 1.4 Coverage

#### 1.4.1 Country

Indicates the country or countries covered in the file

South Africa

### 1.4.2 Geographic Coverage

Information on the geographic coverage of the data. Include the total geographic scope of the data, and any additional levels of geographic coding provided in the variables.

uMkhanyakude district in northern KwaZulu-Natal

### 1.4.3 Universe

A description of the population covered by the data in the file; the group of persons or other elements that are the object of the study and to which the study results refer. Age, nationality, and residence commonly help to delineate a given universe, but any of a number of factors may be involved, such as age limits, sex, marital status, race, ethnic group, etc. The universe may consist of elements other than persons, such as specimen, sample or isolate. In general, it should be possible to tell from the description of the universe whether a given individual or element (hypothetical or real) is a member of the population under study. Also known as universe of interest, population of interest, and target population.

Individuals aged 15 years and above who are resident members of households in the AHRI DSS area

## 1.5 Producers and Sponsors

### 1.5.1 Investigators

The person, corporate body, or agency responsible for the data collection's substantive and intellectual content. Repeat the element for each author and use the affiliation attribute if available. Invert first and last name and use commas. Remarks: The author in this element should be the individual(s) or organization(s) directly responsible for the intellectual content of the data collection.

| Name | Affiliation |
|---|---|
| Professor Deenan Pillay | UCL; AHRI |
| Dr Olivier Koole | LSHTM; AHRI |
| Dr. Emily Wong | AHRI |

### 1.5.2 Funding

The source(s) of funds for production of the data collection. If different funding agencies sponsored different stages of the production process, use the role attribute to distinguish them. Also includes a field for the grant/contract number of the project that sponsored the data collection effort.

| Agency | Abbreviation | Grant number | Role |
|---|---|---|---|
| Wellcome Trust | WT | 097410/Z/11/Z | Core funding |

### 1.5.3 Acknowledgements

Statements of responsibility not recorded in the title and statement of responsibility areas. Indicate here the persons or bodies connected with the work, or significant persons or bodies connected with previous editions and not already named in the description. For example, the name of the person who cleaned the data collection might be cited here, using the role and affiliation attributes. Does not include funders.

| Name | Affiliation | Role |
|---|---|---|
| Dr. Kobus Herbst | AHRI | Chief Information Officer |
| Dickman Gareta | AHRI | Head of Research Data Management |
| Jaco Dreyer | AHRI | Data management, data cleaning |
| Tumi Madolo | AHRI | Data management, data cleaning |
| Siyabonga Nxumalo | AHRI | Data management, data cleaning |

## 1.6 Sampling

### 1.6.1 Sampling Procedure

The type of sample and sample design used to select the survey respondents to represent the population. May include reference to the target sample size and the sampling fraction

N/A. All individuals aged 15 years and older who are resident members of households in the DSS area are eligible to participate Vukuzazi study.

## 1.7 Data Collection

### 1.7.1 Dates of Collection

Contains the date(s) when the data were collected/produced.

Vukuzazi data – 2018

Chest x-ray images

### 1.7.2 Notes on Data Collection/Production

Used to describe noteworthy aspects of the data collection/production.

All resident members aged ≥15 years are eligible to participate in Vukuzazi ("Wake up and know ourselves!" in isiZulu), a platform offering community-based screening for prevalent infections, NCDs and their risk factors using modified WHO-STEPS questionnaires, body measurements, digital chest x-rays and laboratory tests (HIV ELISA, HIV viral load, CD4 cell count, HbA1c and sputum Mtb GeneXpert and culture). Population prevalence of HIV infection, lifetime TB disease, hypertension, diabetes, nutritional status (underweight and obesity) and tobacco and alcohol use for the pre-planned pilot phase were calculated, weighted for nonresponse. Enrollment is ongoing.

## 1.8 Data Processing

### 1.8.1 Other Processing

Used to indicate additional information about the methodology and processing involved in the production of the dataset.

NA

## 1.9 Data Access

We will add these bits

## 1.10 Contacts

### 1.10.1 Contact persons

Individuals listed as contact persons will be used as resource persons regarding problems or questions raised by the user community. The URI attribute should be used to indicate a URN or URL for the homepage of the contact individual. The email attribute is used to indicate an email address for the contact individual.

| Name | Affiliation | Email | URI |
|---|---|---|---|
| Emily Wong | AHRI | emily.wong@ahri.org | |

# 2 File Description

## 2.1 Data Files

### 2.1.1 Contents

Abstract or description of the file. A summary describing the purpose, nature, and scope of the data file, special characteristics of its contents, major subject areas covered, and what questions the PIs attempted to answer when they created the file. No need to repeat information already contained in the abstract in paragraph 1.2.1.

The data contains two files. Image file and clinical data file. The image file contains all images of individuals who participated in the pilot phase of Vukuzazi and had chest x-ray done. The clinical data file of these individuals is also included. The purpose of the data is for devising a machine learning tool to improve upon CAD4TB's triaging of chest x-rays.

# 3 Variable Description

A code book of the variables in the data file.

| Name | Definition | Data Type and Codes |
|---|---|---|
| | | |