1. Title of Dataset: Accuracy of computer aided detection of occupational lung disease: silicosis and pulmonary tuberculosis in ex-miners from the South African gold mines

2. Author Information

A. Principal Investigator Contact Information
Name: Rodney Ehrlich
Institution: University of Cape Town, South Africa
Email: rodney.ehrlich@uct.ac.za

B. Data Analyst Contact Information
Name: Stephen Barker
Institution: University of British Columbia, Canada
Email: stephen.barker@ubc.ca

3. Date of data collection (single date, range, approximate date) <suggested format YYYY-MM-DD>:
Radiographic chest images were collected from 2019-11-25 to 2020-02-13.
External readings were completed from 2020-11-01 to 2021-03-31.
CAD system readings were completed from 2020-03-03 to 2021-03-08.

4. Geographic location of data collection:
South Africa

==========

DATA & FILE OVERVIEW

1. File List:
Dataset-2022.03.18.xlsx

2. Relationship between files, if important:
None

3. Additional related data collected that was not included in the current data package:
None

4. Are there multiple versions of the dataset? yes/no
No

==========

DATA-SPECIFIC INFORMATION FOR: Dataset-2022.03.18.xlsx

1. Number of variables:
25 variables

2. Number of cases/rows:
501 cases, each read by 2 distinct readers (1002 row observations)

3. Variable List:

a. Index: A unique identifier for each of the 501 radiographic chest images. Numeric, range 1-501.
b. Reader: Each chest image was read by 2 human readers. This value (either 1 or 2) indicates the reader.
c. Image quality (score): The quality of the chest image, as determined by the human reader. Range 1 to 4.
d. Image quality (description): The quality of the chest image, as determined by the human reader.
Values correspond to previous column, with 1 = Good (all criteria met), 2 = Good (few technical defects), 3 = Poor (has defects, but can classify), 4 = Unreadable.
e. Profusion (ILO Classification): [Major]/[Minor] profusion as determined by the human reader.
In order, possible values include 0/0, 0/1, 1/0, 1/1, 1/2, 2/1, 2/2, 2/3, 3/2, 3/3.
For more information, see https://www.ilo.org/global/topics/safety-and-health-at-work/resources-library/publications/WCMS_168260/lang--en/index.htm
f. Profusion >= 1/0: Calculated from profusion value, whether the profusion is at least 1/0. 1 if true, 0 if false.
g. Profusion >= 1/1: Calculated from profusion value, whether the profusion is at least 1/1. 1 if true, 0 if false.
h. Profusion >= 2/1: Calculated from profusion value, whether the profusion is at least 2/1. 1 if true, 0 if false.
i. Tuberculosis (TB): Whether the human reader indicates that TB is present in the chest image. 1 if true, 0 if false.
j. TB (Active): The opinion of the human reader as to whether there is any active TB present in the chest image. Possible values are: nil, possible, probable.
k. TB (Prior): The opinion of the human reader as to whether there is any prior/old TB present in the chest image. Possible values are: nil, possible, probable.
l. TB (possible, probable or definite): 1 if the human reader indicated possible/probable/definite active and/or prior TB, 0 otherwise.
m. TB (probable or definite): 1 if the human reader indicated probable/definite active and/or prior TB, 0 otherwise.
n. SilicoTB (profusion >= 1/0): Calculated from previous fields, whether the image is read as having both TB = 1 and profusion at least 1/0. 1 if true, 0 if false.
o. SilicoTB (profusion >= 1/1): Calculated from previous fields, whether the image is read as having both TB = 1 and profusion at least 1/1. 1 if true, 0 if false.
p. Other finding: Other findings read by the human reader.
q. Any abnormality (including profusion 0/1): If the human reader found any abnormality, then 1 (true) otherwise 0 (false).
r. TB and/or profusion >= 1/0: If the human reader identified either or both TB and profusion at least 1/0, then 1 (true) otherwise 0 (false).
s. Worker age (years): the age (in years) of the worker at the time the radiographic chest image was created
t. Worker years of service: the number of years worked. Incomplete, only available in 50% of cases.
u. CAD System A: Abnormal: The score (0-100) that CAD system A has assigned to the image, when detecting abnormality*
v. CAD System A: TB: The score (0-100) that CAD system A has assigned to the image, when detecting TB
w. CAD System A: Silicosis: The score (0-100) that CAD system A has assigned to the image, when detecting silicosis
x. CAD System A: SilicoTB: The score (0-100) that is the average of the TB and silicosis score assigned by System A.
y. CAD System B: Abnormal: The score (0-100) that CAD system B has assigned to the image, when detecting abnormality*
z. CAD System B: TB: The score (0-100) that CAD system B has assigned to the image, when detecting TB
aa. CAD System C: Abnormal: The score (0-100) that CAD system C has assigned to the image, when detecting abnormality*
ab. CAD System C: TB: The score (0-100) that CAD system C has assigned to the image, when detecting TB

* Please see journal publication for a larger discussion of abnormality between the 3 systems.

4. Missing data codes:
Missing fields are left empty

5. Specialized formats or other abbreviations used:
TB = tuberculosis
SilicoTB = silicotuberculosis
ILO = International Labour Organization (https://ilo.org)
CAD = computer aided detection

==========