

# South Africa - Computer Aided Detection of Tuberculosis and Silicosis 2019-2021

**Rodney Ehrlich, Stephen Barker, Jim te Water Naude, David Rees, Barry  
Kistnasamy, Julian Naidoo, Annalee Yassi**

Report generated on: May 15, 2024

Visit our data catalog at: <https://www.datafirst.uct.ac.za/dataportal/index.php>

## Identification

### SURVEY ID NUMBER

zaf-uctfhs-cadts-2019-2021-v1

### TITLE

Computer Aided Detection of Tuberculosis and Silicosis 2019-2021

### COUNTRY

Name	Country code
South Africa	zaf

### STUDY TYPE

Replication/Reproducibility Data

### ABSTRACT

The data was used for the following research publication:

Accuracy of computer aided detection of occupational lung disease: silicosis and pulmonary tuberculosis in ex-miners from the South African gold mines by Rodney Ehrlich (Division of Occupational Medicine at the School of Public Health and Family Medicine, University of Cape Town, South Africa, and Stephen Barker (University of British Columbia, Canada).

The data includes information from radiographic chest images collected from 2019-11-25 to 2020-02-13, external readings from 2020-11-01 to 2021-03-31 and CAD system readings from 2020-03-03 to 2021-03-08.

### KIND OF DATA

Sensor Monitoring Study

## Version

### VERSION DESCRIPTION

v1: Edited, anonymised data for public distribution

### VERSION DATE

2021

## Scope

### NOTES

The data consists of radiographic chest images, external readings, and CAD system readings.

### KEYWORDS

Keyword
Tuberculosis
Silicosis

## Producers and sponsors

### PRIMARY INVESTIGATORS

Name	Affiliation
Rodney Ehrlich	University of Cape Town
Stephen Barker	University of British Columbia

Jim te Water Naude	Diagnostic Research
David Rees	University of the Witwatersrand
Barry Kistnasamy	Office of the Compensation Commissioner for Occupational Diseases
Julian Naidoo	University of the Witwatersrand
Annalee Yassi	University of British Columbia

## Data Collection

### DATES OF DATA COLLECTION

Start	End	Cycle
2019-11-25	2020-02-13	Radiographic chest images
2020-11-01	2021-03-31	External readings
2020-03-03	2021-03-08	CAD system readings

### DATA COLLECTION MODE

Other

## Access policy

### ACCESS CONDITIONS

Public access data for use under a Creative Commons CC-BY (Attribution-only) License

### CITATION REQUIREMENTS

Ehrlich, Rodney, et al. 2022. Computer aided detection of Tuberculosis and Silicosis 2019-2021 [dataset]. Version 1. Cape Town: Datafirst [distributor], 2022. DOI: <https://doi.org/10.25828/sntw-6721>

### ACCESS AUTHORITY

Name	Affiliation	Email	URL
DataFirst	University of Cape Town	support@data1st.org	support.data1st.org

## Metadata production

### PRODUCERS

Name	Affiliation	Role
DataFirst	University of Cape Town	Metadata producer

### DATE OF METADATA PRODUCTION

2024-05-15

### DDI DOCUMENT VERSION

v3

**Data Description**

<b>Data file</b>	<b>Cases</b>	<b>Variables</b>
<b>cadold-2019-2021</b>	1002	28



**Data file: cadold-2019-2021**

Cases: 1002

Variables: 28

**Variables**

ID	Name	Label	Question
V1	Index		
V2	Reader		
V3	Image_quality_score_		
V4	Image_quality_description_		
V5	Profusion_ILO_Classification_		
V6	Profusion___1_0		
V7	Profusion___1_1		
V8	Profusion___2_1		
V9	Tuberculosis_TB_		
V10	TB_Active_		
V11	TB_Prior_		
V12	TB_possible_probable_or_defini		
V13	TB_probable_or_definite_		
V14	SilicoTB_profusion___1_0_		
V15	SilicoTB_profusion___1_1_		
V16	Other_finding		
V17	Any_abnormality_including_profu		
V18	TB_and_or_profusion___1_0		
V19	Worker_age_years_		
V20	Worker_years_of_service		
V21	CAD_System_A_Abnormal		
V22	CAD_System_A_TB		
V23	CAD_System_A_Silicosis		
V24	CAD_System_A_SilicoTB		
V25	CAD_System_B_Abnormal		
V26	CAD_System_B_TB		
V27	CAD_System_C_Abnormal		
V28	CAD_System_C_TB		

Total: 28



**INDEX:****Data file: cadold-2019-2021****Overview**

Valid: 1002   Invalid: 0   Minimum: 1   Maximum: 501   Mean: 251   Standard deviation: 144.698  
 Type: Continuous   Decimal: 0   Width: 3   Range: 1 - 501   Format: Numeric

**TB\_\_ACTIVE\_:****Data file: cadold-2019-2021****Overview**

Valid: 1002   Invalid: 0  
 Type: Discrete   Width: 8   Range: -   Format: character

**Questions and instructions**

## CATEGORIES

Value	Category	Cases	
Nil		909	90.7%
Possible		82	8.2%
Probable		11	1.1%

**READER:****Data file: cadold-2019-2021****Overview**

Valid: 1002   Invalid: 0   Minimum: 1   Maximum: 2   Mean: 1.5   Standard deviation: 0.5  
 Type: Discrete   Decimal: 0   Width: 1   Range: 1 - 2   Format: Numeric

**Questions and instructions**

## CATEGORIES

Value	Category	Cases	
1		501	50%
2		501	50%

**IMAGE\_QUALITY\_SCORE\_:****Data file: cadold-2019-2021****Overview**

Valid: 1002   Invalid: 0  
 Type: Discrete   Width: 1   Range: -   Format: character



## Questions and instructions

### CATEGORIES

Value	Category	Cases	
1		595	59.4%
2		390	38.9%
3		17	1.7%

## IMAGE\_QUALITY\_DESCRIPTION\_:

Data file: cadold-2019-2021

### Overview

Valid: 1002 Invalid: 0  
 Type: Discrete Width: 35 Range: - Format: character

## Questions and instructions

### CATEGORIES

Value	Category	Cases	
Good (all criteria met)		595	59.4%
Good (few technical defects)		390	38.9%
Poor (has defects but can classify)		17	1.7%

## PROFUSION\_ILO\_CLASSIFICATION\_:

Data file: cadold-2019-2021

### Overview

Valid: 1002 Invalid: 0  
 Type: Discrete Width: 4 Range: - Format: character

## Questions and instructions

### CATEGORIES

Value	Category	Cases	
0/0		643	64.2%
0/1		131	13.1%
1/0		36	3.6%
1/1		78	7.8%
1/2		20	2%
2/1		27	2.7%

2/2		54	5.4%
2/3		4	0.4%
3/2		3	0.3%
3/3+		6	0.6%

**PROFUSION\_\_1\_0:****Data file: cadold-2019-2021****Overview**

Valid: 1002   Invalid: 0   Minimum: 0   Maximum: 1   Mean: 0.228   Standard deviation: 0.419  
 Type: Discrete   Decimal: 0   Width: 1   Range: 0 - 1   Format: Numeric

**Questions and instructions**

## CATEGORIES

Value	Category	Cases	
0		774	77.2%
1		228	22.8%

**PROFUSION\_\_1\_1:****Data file: cadold-2019-2021****Overview**

Valid: 1002   Invalid: 0   Minimum: 0   Maximum: 1   Mean: 0.192   Standard deviation: 0.394  
 Type: Discrete   Decimal: 0   Width: 1   Range: 0 - 1   Format: Numeric

**Questions and instructions**

## CATEGORIES

Value	Category	Cases	
0		810	80.8%
1		192	19.2%

**PROFUSION\_\_2\_1:****Data file: cadold-2019-2021****Overview**

Valid: 1002   Invalid: 0   Minimum: 0   Maximum: 1   Mean: 0.0938   Standard deviation: 0.292  
 Type: Discrete   Decimal: 0   Width: 1   Range: 0 - 1   Format: Numeric

## Questions and instructions

### CATEGORIES

Value	Category	Cases	
0		908	90.6%
1		94	9.4%

## TUBERCULOSIS\_TB\_:

Data file: cadold-2019-2021

### Overview

Valid: 1002 Invalid: 0 Minimum: 0 Maximum: 1 Mean: 0.476 Standard deviation: 0.5  
 Type: Discrete Decimal: 0 Width: 1 Range: 0 - 1 Format: Numeric

## Questions and instructions

### CATEGORIES

Value	Category	Cases	
0		525	52.4%
1		477	47.6%

## TB\_PRIOR\_:

Data file: cadold-2019-2021

### Overview

Valid: 1002 Invalid: 0  
 Type: Discrete Width: 8 Range: - Format: character

## Questions and instructions

### CATEGORIES

Value	Category	Cases	
Nil		528	52.7%
Possible		133	13.3%
Probable		341	34%

## TB\_POSSIBLE\_PROBABLE\_OR\_DEFINI:

Data file: cadold-2019-2021

**Overview**

Valid: 1002   Invalid: 0   Minimum: 0   Maximum: 1   Mean: 0.476   Standard deviation: 0.5  
 Type: Discrete   Decimal: 0   Width: 1   Range: 0 - 1   Format: Numeric

**Questions and instructions**

## CATEGORIES

Value	Category	Cases	
0		525	52.4%
1		477	47.6%

**TB\_PROBABLE\_OR\_DEFINITE\_:**

Data file: cadold-2019-2021

**Overview**

Valid: 1002   Invalid: 0   Minimum: 0   Maximum: 1   Mean: 0.343   Standard deviation: 0.475  
 Type: Discrete   Decimal: 0   Width: 1   Range: 0 - 1   Format: Numeric

**Questions and instructions**

## CATEGORIES

Value	Category	Cases	
0		658	65.7%
1		344	34.3%

**SILICOTB\_PROFUSION\_\_1\_0\_:**

Data file: cadold-2019-2021

**Overview**

Valid: 1002   Invalid: 0   Minimum: 0   Maximum: 1   Mean: 0.14   Standard deviation: 0.347  
 Type: Discrete   Decimal: 0   Width: 1   Range: 0 - 1   Format: Numeric

**Questions and instructions**

## CATEGORIES

Value	Category	Cases	
0		862	86%
1		140	14%

**SILICOTB\_PROFUSION\_\_1\_1\_:****Data file: cadold-2019-2021****Overview**

Valid: 1002   Invalid: 0   Minimum: 0   Maximum: 1   Mean: 0.119   Standard deviation: 0.324  
 Type: Discrete   Decimal: 0   Width: 1   Range: 0 - 1   Format: Numeric

**Questions and instructions**

## CATEGORIES

Value	Category	Cases	
0		883	88.1%
1		119	11.9%

**OTHER\_FINDING:****Data file: cadold-2019-2021****Overview**

Valid: 261   Invalid: 0  
 Type: Discrete   Width: 50   Range: -   Format: character

**Questions and instructions**

## CATEGORIES

Value	Category	Cases	
Bullae		20	7.7%
Bullae, Hyperinflation		5	1.9%
Bullae, Progressive massive fibrosis (PMF)		1	0.4%
Cardiomegaly		36	13.8%
Hyperinflation		173	66.3%
Hyperinflation, Cardiomegaly		2	0.8%
Hyperinflation, Progressive massive fibrosis (PMF)		5	1.9%
Other infection		3	1.1%
Pneumothorax		1	0.4%
Pneumothorax, Bullae		1	0.4%
Progressive massive fibrosis (PMF)		6	2.3%
Solid mass		4	1.5%
Solid mass, Trauma including fractures		1	0.4%
Trauma including fractures		3	1.1%

**ANY\_ABNORMALITY\_INCLUDING\_PROFU:****Data file: cadold-2019-2021****Overview**

Valid: 1002   Invalid: 0   Minimum: 0   Maximum: 1   Mean: 0.813   Standard deviation: 0.39  
 Type: Discrete   Decimal: 0   Width: 1   Range: 0 - 1   Format: Numeric

**Questions and instructions**

## CATEGORIES

Value	Category	Cases	
0		187	18.7%
1		815	81.3%

**TB\_AND\_OR\_PROFUSION\_\_1\_0:****Data file: cadold-2019-2021****Overview**

Valid: 1002   Invalid: 0   Minimum: 0   Maximum: 1   Mean: 0.564   Standard deviation: 0.496  
 Type: Discrete   Decimal: 0   Width: 1   Range: 0 - 1   Format: Numeric

**Questions and instructions**

## CATEGORIES

Value	Category	Cases	
0		437	43.6%
1		565	56.4%

**WORKER\_AGE\_YEARS\_:****Data file: cadold-2019-2021****Overview**

Valid: 994   Invalid: 8   Minimum: 32   Maximum: 92   Mean: 61.98   Standard deviation: 7.665  
 Type: Continuous   Decimal: 0   Width: 2   Range: 32 - 92   Format: Numeric

**WORKER\_YEARS\_OF\_SERVICE:****Data file: cadold-2019-2021****Overview**

Valid: 514   Invalid: 488   Minimum: 0.92   Maximum: 76.95   Mean: 21.703   Standard deviation: 10.663  
 Type: Continuous   Decimal: 2   Width: 5   Range: 0.92 - 76.95   Format: Numeric

**CAD\_SYSTEM\_A\_\_ABNORMAL:****Data file: cadold-2019-2021****Overview**

Valid: 994   Invalid: 8   Minimum: 2   Maximum: 98   Mean: 47.543   Standard deviation: 30.305  
Type: Continuous   Decimal: 0   Width: 2   Range: 2 - 98   Format: Numeric

---

**CAD\_SYSTEM\_A\_\_TB:****Data file: cadold-2019-2021****Overview**

Valid: 994   Invalid: 8   Minimum: 1   Maximum: 96   Mean: 44.26   Standard deviation: 29.268  
Type: Continuous   Decimal: 0   Width: 2   Range: 1 - 96   Format: Numeric

---

**CAD\_SYSTEM\_A\_\_SILICOSIS:****Data file: cadold-2019-2021****Overview**

Valid: 994   Invalid: 8   Minimum: 1   Maximum: 98   Mean: 30.803   Standard deviation: 29.906  
Type: Continuous   Decimal: 0   Width: 2   Range: 1 - 98   Format: Numeric

---

**CAD\_SYSTEM\_A\_\_SILICOTB:****Data file: cadold-2019-2021****Overview**

Valid: 994   Invalid: 8   Minimum: 1.5   Maximum: 94   Mean: 37.531   Standard deviation: 27.348  
Type: Continuous   Decimal: 2   Width: 5   Range: 1.5 - 94   Format: Numeric

---

**CAD\_SYSTEM\_B\_\_ABNORMAL:****Data file: cadold-2019-2021****Overview**

Valid: 1002   Invalid: 0   Minimum: 1   Maximum: 99   Mean: 78.713   Standard deviation: 26.022  
Type: Continuous   Decimal: 0   Width: 2   Range: 1 - 99   Format: Numeric

---

**CAD\_SYSTEM\_B\_\_TB:****Data file: cadold-2019-2021**

## Overview

Valid: 1002   Invalid: 0   Minimum: 0   Maximum: 98   Mean: 43.257   Standard deviation: 38.887  
Type: Continuous   Decimal: 0   Width: 2   Range: 0 - 98   Format: Numeric

---

### CAD\_SYSTEM\_C\_\_ABNORMAL:

Data file: cadold-2019-2021

## Overview

Valid: 1002   Invalid: 0   Minimum: 0.26   Maximum: 98.328   Mean: 24.657   Standard deviation: 26.265  
Type: Continuous   Decimal: 2   Width: 5   Range: 0.260001420974698 - 98.32822494   Format: Numeric

---

### CAD\_SYSTEM\_C\_\_TB:

Data file: cadold-2019-2021

## Overview

Valid: 1002   Invalid: 0   Minimum: 0.13   Maximum: 98.328   Mean: 21.09   Standard deviation: 26.747  
Type: Continuous   Decimal: 2   Width: 5   Range: 0.130000000353903 - 98.3282268   Format: Numeric

---



## Download related resources

### Other materials

#### ehrllich-barker-data-readme

---

Title	ehrllich-barker-data-readme
Author(s)	Rodney Ehrlich and Stephen Barker
Date	2022-01-01
Country	South Africa
Language	English
Contributor(s)	This documents explains the contents of the dataset and the related research publication
Filename	ehrllich-barker-data-readme.pdf

---